

UNITED STATES AIR FORCE RESEARCH LABORATORY



STUDIES AND ANALYSES OF AIDED ADVERSARIAL DECISION MAKING PHASE 2: RESEARCH ON HUMAN TRUST IN AUTOMATION

James Llinas
Ann Bisantz
Colin Drury
Younho Seong
Jiun-Yin Jian

STATE UNIVERSITY OF NEW YORK AT BUFFALO
CENTER OF MULTISOURCE INFORMATION FUSION
DEPARTMENT OF INDUSTRIAL ENGINEERING
BUFFALO NY 14260-2050

APRIL 1998

INTERIM REPORT FOR THE PERIOD 10 APRIL 1997 TO 31 MAY 1998

Approved for public release; distribution is unlimited.

Human Effectiveness Directorate
Crew System Interface Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

19991130 027

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Air Force Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, Virginia 22060-6218

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-1999-0216

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER



MARIS M. VIKMANIS
Chief, Crew System Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE April 1998	3. REPORT TYPE AND DATES COVERED Interim Report; 10 April 97 to 31 May 98		
4. TITLE AND SUBTITLE Studies and Analyses of Aided Adversarial Decision Making. Phase 2: Research on Human Trust in Automation		5. FUNDING NUMBERS C: F41624-94-D-6000 PE: 62202F PR: 7184 TA: 10 WU: 46		
6. AUTHOR(S) James Llinas, Ann Bisantz, Colin Drury, Younho Seong, & Jiun-Yin Jian				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) State University of New York at Buffalo Center of Multisource Information Fusion Department of Industrial Engineering Buffalo, NY 14260-2050		8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory, Human Effectiveness Directorate Crew System Interface Division Air Force Materiel Command Wright-Patterson AFB, OH 45433-7022		10. SPONSORING/MONITORING AGENCY REPORT NUMBER AFRL-HE-WP-TR-1999-0216		
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.		12b. DISTRIBUTION CODE		
13. ABSTRACT (Maximum 200 words) This report describes the second phase of work conducted at the Center for Multi-source Information Fusion at the State University of New York at Buffalo. This work focused on Aided Adversarial Decision Making (AADM) in Information Warfare (IW) environments. Previous work examined informational dependencies and vulnerabilities in AADM to offensive IW operations. In particular, human trust in automated, information warfare environments was identified as a factor which may contribute to these vulnerabilities and dependencies. Given that offensive IW operations may interfere with automated, data-fusion based decision aids, it is necessary to understand how personnel may rely on or trust these aids when appropriate (e.g., when the information provided by the aids is sound), and recognize the need to seek other information (i.e., to "distrust" the aid) when the information system has been attacked. To address these questions, this report details background research in the areas of human trust in automated systems and sociological findings on human trust, details the development of an empirically-based scale to measure trust, provides a framework for investigating issues of human trust and its effect on performance in an AADM-IW environment, and describes the requirements for a laboratory designed to conduct these investigations.				
14. SUBJECT TERMS decision making, decision aids, decision models, adversarial decision making, data fusion, trust in automation, human-machine interaction, information warfare		15. NUMBER OF PAGES 117		
		16. PRICE CODE		
17. SECURITY CLASSIFICATION OF REPORT UNCLASSIFIED	18. SECURITY CLASSIFICATION OF THIS PAGE UNCLASSIFIED	19. SECURITY CLASSIFICATION OF ABSTRACT UNCLASSIFIED	20. LIMITATION OF ABSTRACT UNL	

This page intentionally left blank.

PREFACE

This effort was accomplished under Contract F41624-94-D-6000, Delivery Order 0007 for the Air Force Research Laboratory's Human Effectiveness Directorate, under the direction of the Crew System Interface Division, Information Analysis and Exploitation Branch (AFRL/HECA). It was completed for the prime contractor, Logicon Technical Services, Inc. (LTSI), Dayton, Ohio, under Work Unit No. 71841046: "Crew Systems for Information Warfare." Mr. Don Monk was the Contract Monitor.

The authors offer special thanks to Mr. Gilbert Kuperman, of AFRL/HECA, the Work Unit Manager, for his initial interest, and for his ongoing support and direction which made this effort possible. The authors also wish to acknowledge Mr. Robert L. Stewart of LTSI, for his management of the project, and Ms. Elisabeth Fitzhugh, also of LTSI, for technical editing services. The Semi-Automated Ground Environment (SAGE) was made available by Ball Aerospace Corp., whose support is gratefully acknowledged.

TABLE OF CONTENTS

LIST OF FIGURES.....	vi
LIST OF TABLES	vii
1.0 SETTING THE PERSPECTIVE: AIDED ADVERSARIAL DECISION MAKING AND INFORMATION WARFARE	1
1.1 Motivating Problem Framework—a Model	3
1.1.1 Assumptions and Constraints.....	3
1.1.2 Two-Sided Adversarial General Model	3
1.1.3 Aided Human Decision Making Model	6
1.2 Dimensions of the Problem.....	6
1.2.1 Automated Data Fusion as a Decision Aid	6
1.2.2 Informational Value in Decision Making	8
1.2.3 Errors In Human Decision Making	10
1.2.3.1 What is Error?	10
1.2.4 Cultural Effects On Adversarial Decision Making	11
1.2.4.1. Basic Notions of Culture: What Culture Is (Hoecklin, 1995)	12
1.2.5 Trust in Automation.....	12
1.2.5.1 State of Research	13
1.2.5.2 Behavioral Response to Distrusted Systems	13
1.3 References	13
2.0 CONCEPTS AND DEFINITIONS OF TRUST AND RELATED NOTIONS	16
2.1 Overview of the Sociological Literature	16
2.2 Overview of the Human Factors Engineering Literature	19
2.2.1 Supervisory Control and Automation	19
2.2.2 Models of Trust.....	20
2.3 Characteristics of Trust	25
2.3.1 Characteristics of the “Other Actor”	26
2.3.2 Characteristics of Data Communicated.....	28
2.3.3 Dynamics of Trust: Overtrust-Trust-Distrust-Mistrust	33
2.3.3.1 Calibration of Trust and Mistrust	33
2.3.3.2 Overtrust (Complacency)	35
2.3.3.3 Distrust	38
2.4 Implications for IW	39
2.4.1 What Do We Know About Trust?	39
2.4.2 A Model of Trust for IW.....	39
2.4.3 Applying the Lens Model to Human Trust in Complex, Automated Systems	42
2.4.4 Instantiating the Model	43
2.5 References	45

3.0 MEASURES OF TRUST AND RELATED NOTIONS	52
3.1 Overview	52
3.2 Rating Measures of Trust	52
3.3 Developing an Empirically Based Scale to Measure Trust	55
3.5 Performance and Process Measures	62
3.6 Summary	63
3.7 References	64
4.0 INVESTIGATING TRUST IN AN IW DOMAIN	66
4.1 Introduction	66
4.2 Previous Investigations of Trust in Automated Support Systems	66
4.3 Designing Experimental Scenarios for Studies of Trust in Aided Adversarial Decision Making (AADM) Environments	67
4.3.1 Developing a Framework for Experimentation	67
4.3.1.1 System Dimension	67
4.3.1.2 Surface-Depth Dimension	67
4.3.1.3 Further Categories of Corruption	68
4.3.2 Experimental Context	69
4.3.2.1 Experimental Scenarios and Manipulations	69
4.3.2.2 Experimental Participants	71
4.4 References	71
5.0 IMPLICATIONS FOR THE DESIGN OF AN INFORMATION WARFARE LABORATORY	72
5.1 Overview	72
5.2 Laboratory Design for Pilot Studies	73
5.3 A Specific Laboratory Concept	74
5.4 References	77
GLOSSARY	79
APPENDIX A	80
APPENDIX B	85
APPENDIX C	94
APPENDIX D	101

LIST OF FIGURES

Figure	Page
1.1 Prominent IW Activities	2
1.2 Two-Sided General Model.....	4
1.3 Aided Human Decision Making Model.....	5
1.4 JDL/DFG Data Fusion Process Model.....	7
2.1 Conceptual diagram representing the interactive process display interface between the human operators and the automated system used in Muir and Moray (1996).....	29
2.2 The system properties considered in human-machine interaction can be described at various levels of abstraction, representing the physical implementation and functional purpose in varying degrees.	32
2.3 Two vicious cycles on the trust continuum.....	38
2.4 Trust transmission model.....	41
2.5 Brunswik's Lens Model adapted from Cooksey (1996).	42
2.6 Model of human trust in automation using the Lens model.	45
3.1 Ratings of unlabeled trust vs. distrust, for 112 words.....	58
3.2 Ratings of Human-human trust vs. distrust, for 112 words.	58
3.3 Ratings of Human-machine trust vs. distrust, for 112 words.....	59
3.4 Union set size for words negatively related to trust.	61
3.5 Union set size for words positively related to trust.....	61
5.1 Current state of research on trust and IW.....	72
5.2 Synthetic AADM Model with SAGE	75
5.3 Two-Sided General Model.....	76
5.4 Detailed AADM Lab Concept Using "LENS" Model IW Framework.....	78
C.1 Brunswik's Lens Model	97
C.2 Model of human trust in automation using the Lens model.....	99
D.1 Components of an AADM Environment described along a system and a surface-depth dimension. Potential experimental scenarios and manipulations, organized by system and surface-depth dimensions and levels of malfunction, causal factors.	107
D.2 Experimental concept for AADM IW	108

LIST OF TABLES

Table	Page
1.1 Uncertainty.....	10
1.2 Error Genotypes, Adapted from Reason (1990).....	11
2.1 Association of Barber's Technical Competent Performance to Rasmussen's Taxonomy...	20
2.2 Muir's Framework for Studying Trust in Supervisory Control Environments, Produced by Crossing Barber's (1983) Taxonomy of Trust (Rows) with Rempel, Holmes, and Zanna's (1985) Taxonomy of the Development of Trust (Columns). Adapted from Muir (1989).	21
2.3 Proposed Dimensions and Relationship Between the Different Dimensions of Trust. Adapted from Lee (1992)	24
2.4 Selected Conditions Used in Muir and Moray (1996)'s Experiments	30
2.5 How the Operator's Trust In and Use Of the Automation Interact with the Quality of the Automation to Influence System Performance. Adapted from Muir (1994).	34
3.1 Selected Questions from Rempel et al.'s (1985) Trust Scale..	53
3.2 Example Questions from the Interpersonal Relationship Scale (Rotter, 1967)..	53
3.3 Example Statements from the Dyadic Trust Scale Study (Larzelere & Huston, 1980)..	54
3.4 Example Motivational Statements from the Complacency Potential Rating Scale Study (Singh et al., 1993)..	54
3.5 Motivational Statements from the Subjective Rating Scale Study of Lee and Moray (1994)..	55
3.6 Word Sets Related to Human, Human-Machine, and General Trust.....	60
4.1 Components of an Aided Adversarial Decision Making Environment Described Along a System and a Surface-Depth Dimension.	68
4.2 Potential Experimental Scenarios and Manipulations, Organized by System and Surface-Depth Dimensions and Levels of Malfunction, Causal Factors.	70
A.1 Notions of Trust and Related IW Features.....	83
A.2 Possible Information Operation Attacks on Trust Attributes.....	84
B.1 Social-Psychology Notions of Trust.....	86
B.2 Human Factors Notions of Trust.....	88

This page intentionally left blank.

1.0 SETTING THE PERSPECTIVE: AIDED ADVERSARIAL DECISION MAKING AND INFORMATION WARFARE

This report describes the second phase of work conducted at the Center for Multisource Information Fusion (CMIF) at the State University of New York at Buffalo (SUNY@Buffalo; hereafter "UB") in the research area of Aided Adversarial Decision Making (AADM). In particular, this work has focused on AADM in Information Warfare environments. The first phase (see Llinas, Drury, Bialas & Chen, 1997) examined a number of factors that surround this topic; one particular emphasis was in examining informational dependencies and vulnerabilities in AADM to offensive Information Warfare (IW) operations, but in exploring this issue it was realized that many other factors influence the nature of AADM, and these factors were also examined (e.g., informational value in decision making, cultural differences in AADM, patterns of human error, etc.). Brief comments are made here on this prior work to help set the stage for the material discussed in the body of this Phase 2 report.

The second phase of this work focused in particular on "Human Trust in Automation" in IW environments. While the first phase uncovered many factors that give shape to the overall problem of AADM, in discussions with Air Force Research Laboratory (AFRL) staff it was realized that in spite of the possible (and combinatorial) effects of many of these factors, a root issue is whether users trust the computer-based decision aids¹ they are using. In our literature searches on this topic (described in Section 2), it was somewhat surprising, in the era of the Internet/Web, to see very little work on this subject. Particularly lacking were experimental studies with human subjects. It was expected that some body of work on this topic would have been uncovered; in many modern-day defense applications involving the use of computer-based automation, users frequently ignore, or worse yet, turn off automated support systems for a variety of reasons. Some of these reasons don't have to do with trust explicitly, but many are trust-based, so this issue seems rather fundamental to the effective utilization of workstation-based decision aids of various types. Further, in most prior works, including those we reviewed in the sociological literature, the notion of trust was studied in the framework of *friendly* actors. Our focus is on the *adversarial* case, where situations are created by hostile "Information Operations,"² in which the integrity of the information being processed by the decision aid/data fusion process is suspect. This is in addition to the possibilities for deception, etc. that comprise the field of "counterinformation"; see Figure 1.1 on the next page.

Thus, we see the trust in automation issue as a research topic rich with intellectual issues and having potentially high payoff to the military. If deep understanding about the nature of trust establishment, trust loss, mistrust, distrust, etc., can be developed, improvements in both system design and in operational procedures should be feasible, leading to high payoff in the sense of effective use of decision-aided systems even when under information attack. In this effort, we have striven to provide a solid understanding of the multi-dimensional characteristics of trust,

¹ In this, and in the prior report, the automated decision aid is postulated as an automated data fusion process supporting both individual target position and identification (ID) estimates ("Level 1 fusion"), situational estimates ("Level 2 fusion"), and threat estimates ("Level 3 fusion"). The prior phase report discusses the role of automated data fusion in AADM, and information dependencies and vulnerabilities of the data fusion process.

² This is the term that the IW community seems to have chosen to signify offensive-type IW activities.

and have also centered part of our effort on understanding various aspects (e.g., metrics) necessary to the conduct of well-designed experiments in which aspects of trust in automation can be studied empirically. We hope to both establish a *Laboratory for Information Warfare Studies* and conduct experiments in the lab in our next phase.

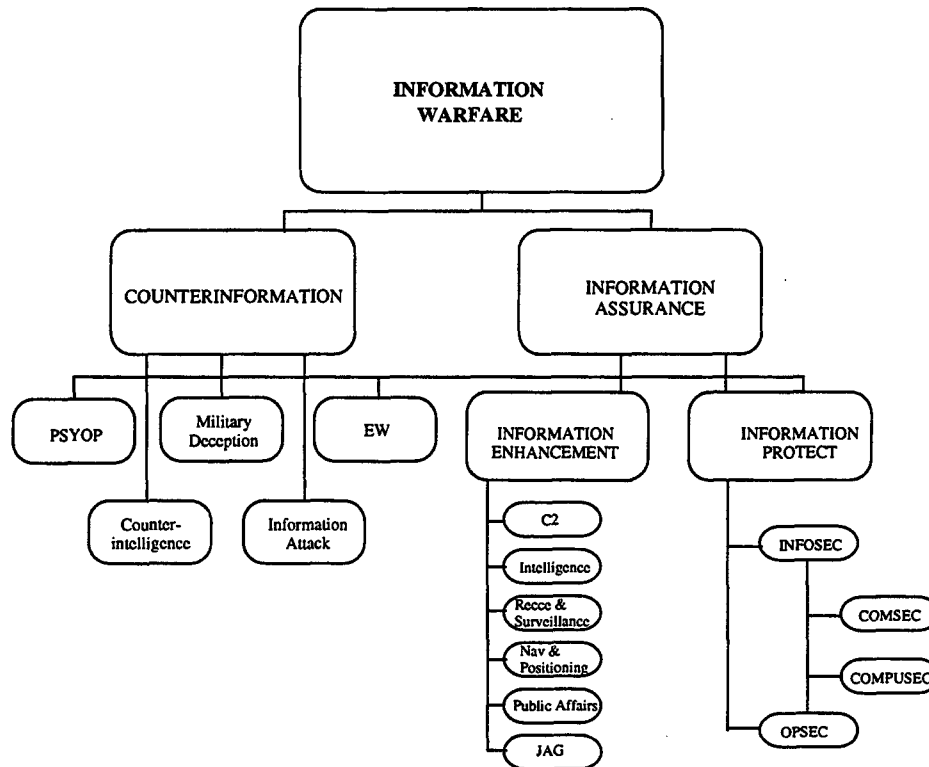


Figure 1.1 Prominent IW Activities

In our work, we have made reference, for the purpose of familiarization, to a sampling of works from the extensive and very dynamic body of literature dealing with the elusive and complex topic of IW³. These works reveal that the subject is quite intricate and complicated, from the policy level down to the operational level. In particular, offensive IW operations turn out to be a sensitive, and apparently classified subject, most generally available works on IW operations are about defensive operations, involving such issues as encryption and procedural security, etc. In our work, we simply postulate environments where the integrity of the information in the aiding system (the computer-based decision aid, or data fusion process), is suspect, due to *whatever* offensive IW techniques (a simple example is viral attack). It is understood that to deal with the subject of trust in a thorough way we will need better understanding of offensive IW techniques, but we focus on the basic issues for now, given the primitive state of understanding of trust in automation.

³ See FM 100-6, 1994; Luoma, 1994; Denning, 1990; Szafranski, 1995; Stein, 1995; and Libicki, 1997 for examples of the works on IW

1.1 Motivating Problem Framework—a Model

We mentioned above that IW is a multi-dimensional topic. In our first phase, we attempted to combine these dimensions into an aided decision making model described here. The aided decision model presented in the following has two levels: 1) the Two-Sided Adversarial General Model (Fig 1.2) is the upper level model which describes the relationship between two adversarial forces, both of whom have decision aids and 2) the Aided Human Decision Making Model (Fig 1.3), which is based on the model of situation awareness by Endsley (1995) and which also integrated the concepts from the modified Recognition-Primed Decision (RPD) model by Kaempf, Klein, Thordsen and Wolf (1996) and the Mixed Initiative Model (MIM) by Riley (1989), which focuses on the detailed information processing in the human-decision aided cooperative system of either side of the adversarial forces in the general model.

1.1.1 Assumptions and Constraints

The proposed aided decision making model is focused on the following considerations.

- The decision making tasks discussed in the Phase 1 study (primarily focused on tactical rather than operational decision making).
- A decision environment which is assumed to be adversarial, complex, time-pressured, risky, dynamic, and involving various types of uncertainty.
- A decision making process supported by a data fusion-based decision aid. The information to the decision maker is primarily provided by the display interfaces of the decision aid.
- Cases where the decision makers are assumed to be experienced and well-trained in the designated command and control tasks and in interacting with the decision-aiding system.

The case of the single decision-maker; that is, group or distributed types of aided decision making were not considered in this study.

1.1.2 Two-Sided Adversarial General Model

This general model, as shown in Figure 1.2, depicts the information flow between the two opposing forces—adversary and friendly. For each side, three major nodes are addressed: the human commander (the decision maker), the data fusion system (the decision aid), and the world (as perceived through information resources). As shown in the diagram, in order for human commanders to perform command and control of the battlefield, the sensors collect data (which often relate to the states of environment, antagonists, and protagonists) from the battlefield or the world and feed this information into the data processing/fusion system; the processed information is displayed to the human commander who can then make a decision. In addition, the supporting information, other than current battlefield information, may also be accessed through those available databases connecting to the decision-aiding system. Decision makers gather information primarily by interacting with the display and control interface(s) provided by

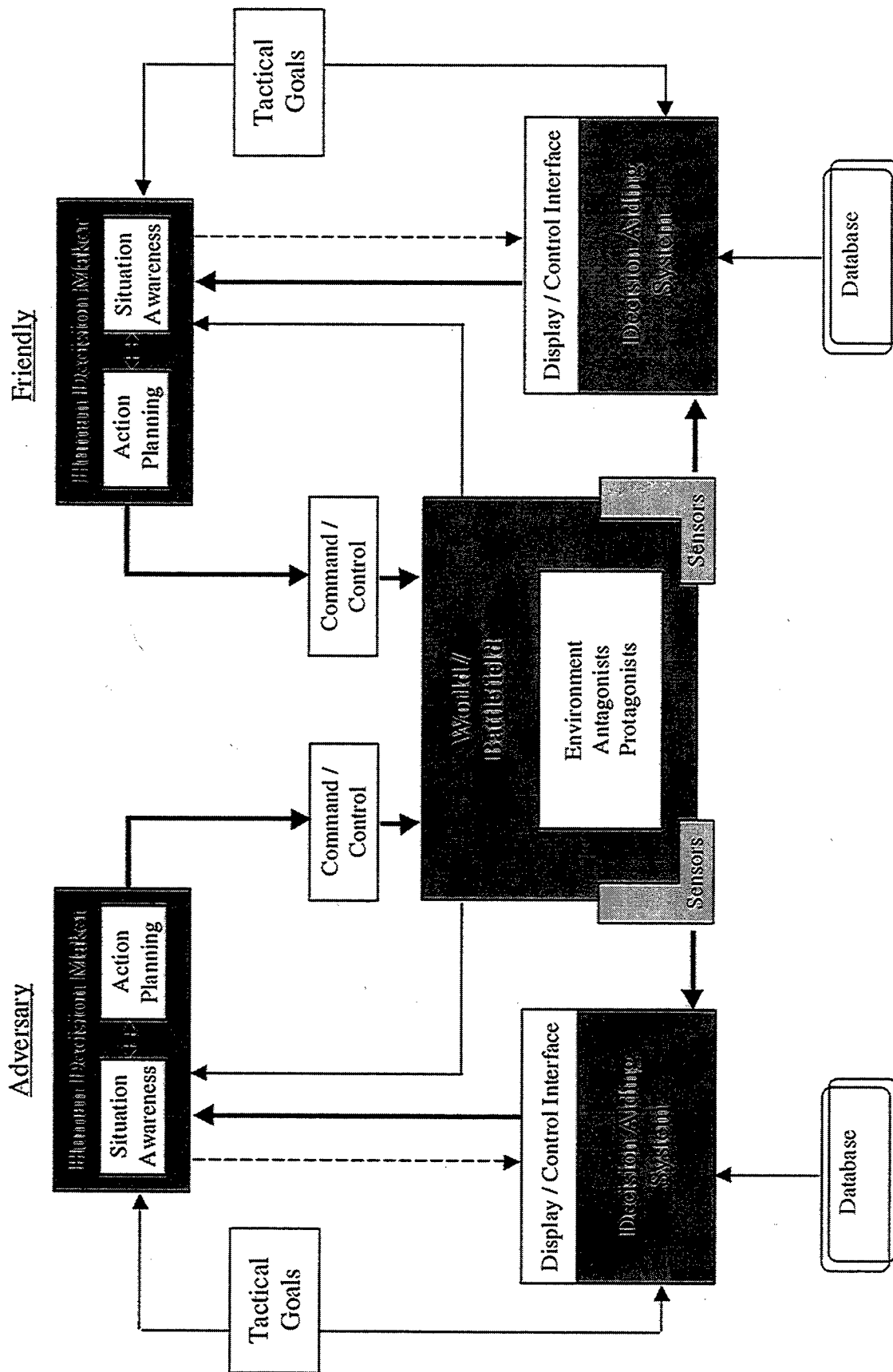


Figure 1.2 Two-Sided General Model

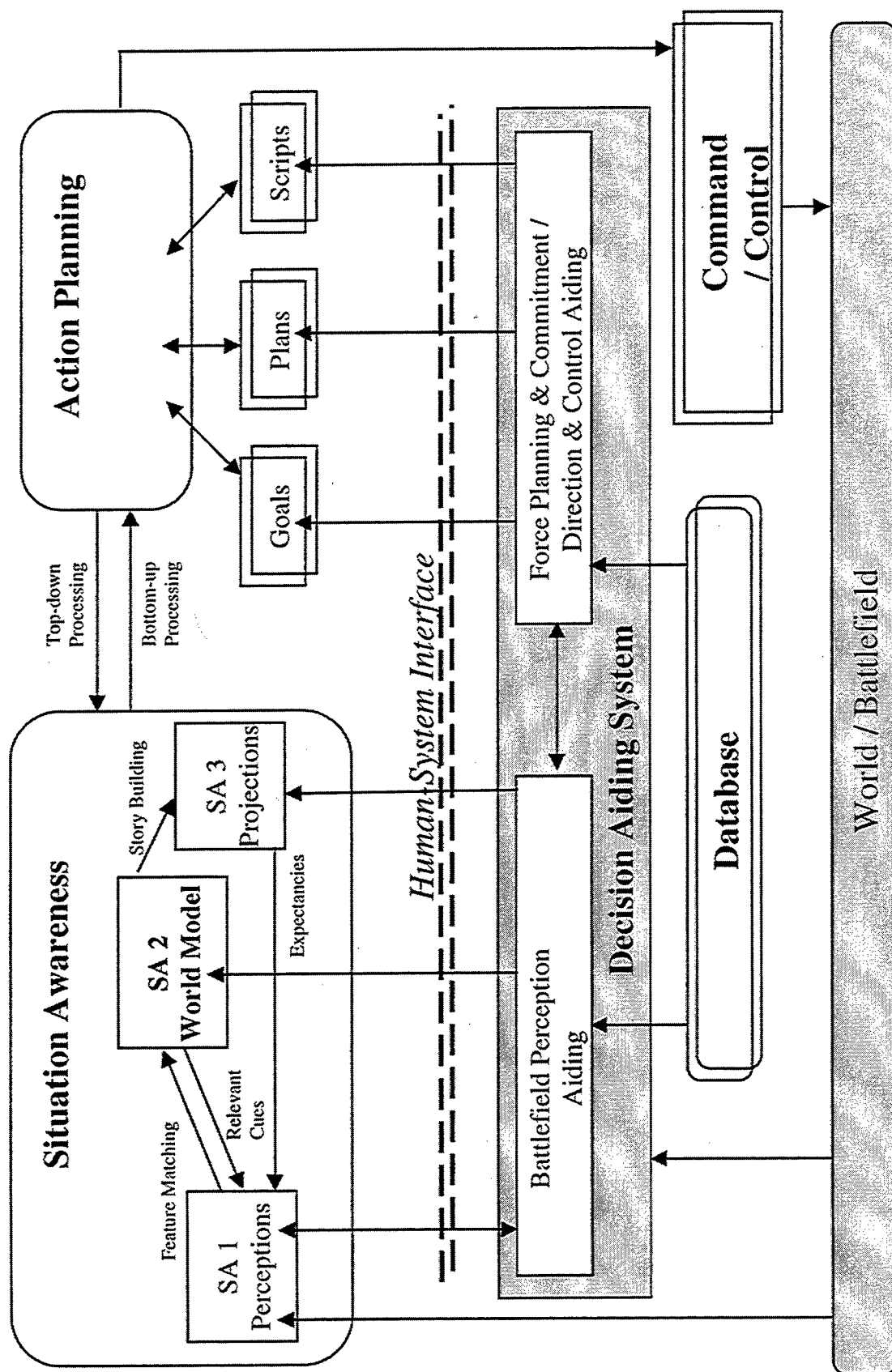


Figure 1.3 Aided Human Decision Making Model

the decision-aiding system. Furthermore, the decision maker may also collect information through other (perhaps unofficial) sources. Using the information gathered, the decision maker thus forms a situation model and, based on the model, plans the command and control actions.

Although the decision making paradigms in this model are generally the same for both sides, the information flows or decision making results can be very different. The differences in characteristics between the entities are the determinants for different decision making results. The information dependencies and vulnerabilities in aided adversarial decision making are determined by the entity characteristics in this model. Technology is one of the major factors, especially the ability to design and use sophisticated data sensors, data processing systems, interactive interfaces, and communications systems for command and control. Cultural differences are another potential factor. Different cultures can result in different command and control patterns, different personal decision making processes, and thus require different decision aids for maximum effectiveness.

1.1.3 Aided Human Decision Making Model

As shown in Figure 1.3 on the page before, this model depicts a rather detailed view inside the decision making process, especially for the human side. The basic relationships among the human decision maker, decision aid, and battlefield have already been defined in the general model described in Section 1.1.2. The human decision making process can be categorized as comprising two major phases: situation awareness and action planning. That is, based upon the information gathered from the decision aid or/and from the world, the decision maker can form a situation model and then, based on the model, plan the command and control actions. This part of the model will be discussed in more detail below.

1.2 Dimensions of the Problem

1.2.1 Automated Data Fusion as a Decision Aid

As noted above, this project deals with *aided* adversarial decision making (DM). The legacy of research in computer-based decision aids is extensive, and for several years there were conferences which focused on and discussed the work being done by a rather large community in the development of not only prototype aids themselves but also on the underlying principles of decision aid development (i.e., both the "what" and the "how to"). In the present case, we hypothesize that the general nature of a computer-based decision aid for each adversary in our general model takes the form of, and is based on, the notion that a *data or information fusion* process provides the basis for the aid. Data fusion (DF) processing is itself a rather broad and multidisciplinary topic but can be modeled, to about the same level of fidelity as our general model described above, by a "process model" originally developed by the Joint Directors of Laboratories Data Fusion Group (JDL/DFG), a defense laboratory data fusion technology oversight committee. This general model is shown in Figure 1.4 on the next page and can be seen to comprise 4 "Levels" of processing, which are discussed in detail below.

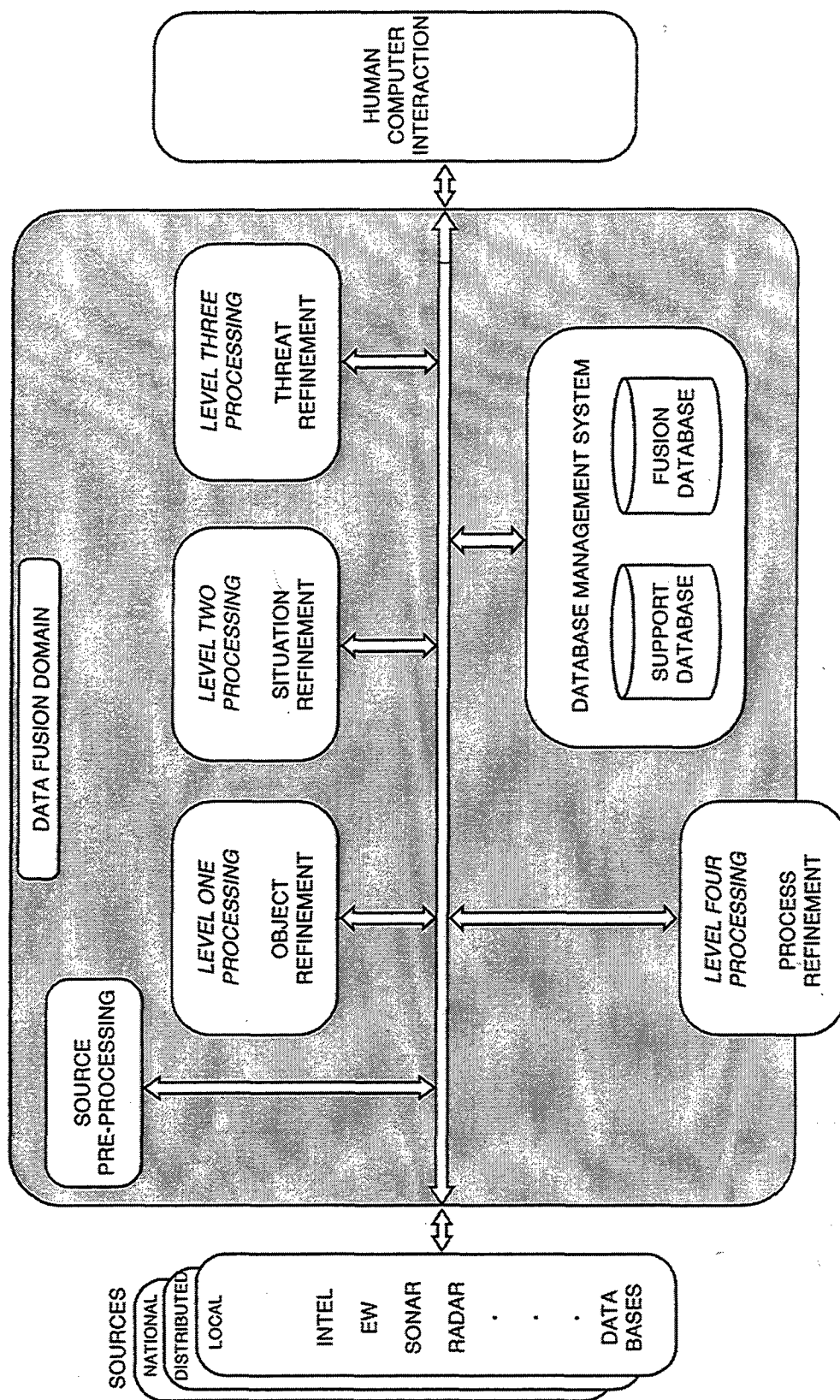


Figure 1.4 JDL/DFG Data Fusion Process Model

Level 1: those processes involved with normalizing a set of multisource-based inputs (a data preparation step prior to combining/fusing), association-correlation-assignment processing (to relate observations to hypothetical objects and related estimation processes about object features), and fusion-based estimation processes which estimate both the kinematics and identity of the targets hypothesized. Thus, Level 1 processing produces what might be called a "labeled set"—of individual targets (points in space), each having labels describing their kinematic and identity properties.

Level 2: those processes involved with situational estimation, traditionally involving aggregation of single (Level 1) object estimates into "order of battle (OB)" structures (i.e., aggregated targets), and behaviorally-based estimates (events and activities). Thus, Level 2 processing, which is largely (but not exclusively) symbolically based, produces a *contextual interpretation* of an abstraction, typically labeled a "situation," by fusing Level 1 estimates, a priori knowledge, and other observations.

Level 3: those processes, also predominantly symbolic in nature, that produce what is in essence a "special" situation estimate traditionally called a "threat" estimate. A threat state or situation is distinguished from benign situations by three factors: the idea that a lethal capability ("*lethality*") exists on the hostile side, that there is an *opportunity* to employ that lethality, and that there is an *intent* to use that lethality. Hence, these processes focus on estimating these factors in particular.

Level 4: those processes which enable a sense of "intelligent control" or adaptation of the overall fusion process. Typically these processes are considered to involve either control over (a) input source/sensor operations (sometimes called "sensor management" or "collection management"), and/or (b) intelligent adaptation of the internal (Levels 1–3) processes of the DF process itself. Control of the latter type can be implemented either parametrically or by controlled switching and optimization in the use of multiple algorithms or processes for a given DF function.

1.2.2 Informational Value in Decision Making

Central to the analysis of the effects of IW on decision making is the assessment of Informational Value in DM. IW attacks will lead to the deletion, corruption, and alteration of quanta of data and/or information in any automated DM support system or in the mind of the user/analyst/operator. Presuming validity in the assertions of the Defense Information Systems Agency (DISA) which argues that perfect protection of information in networks is impossible, or at least unaffordable into the mid-term future, information will indeed be compromised and systems should be designed under this assumption. So the immediate question is, if this happens, "so what?"

The approaches taken in Phase 1 drew heavily from the works of Morris (1964), Yovits and Abilock (1974), and Ackoff (1958), among others, each of whom has examined the question of informational value in DM in somewhat different but related ways. The models and concepts drawn from these references were observed to also have similarities to those from the theories associated with Reinforcement Learning, which also constructs a probabilistic (expectation-centered) model of the roles of information in learning processes.

Much of what was synthesized in this research was excerpted from Yovits and Abilock (1974). Additionally, Morris (1964) pointed out that it is now generally recognized that "information theory" is neither a rival to, nor a substitute for, a general theory of signs (i.e., semiotics). The frequently-cited Shannon and Weaver information theoretic viewpoint concerns the transmission of a message as a symbol string independent of its content. Bar-Hillel (1955) and MacKay (1952) took alternative views. Further, MacKay regarded information as that which changes our representations—that is, our signs. Gaining information is thus a mechanism for changing our expectations (i.e., our dispositions to respond), caused by a sign. He distinguished between *selective* and *semantic* information. Selective information gives the information necessary to select the message itself and is not concerned with the content of the message; it is in some sense a signaling theory. Semantic information, on the other hand, is concerned with the content of the message⁴. Shannon's theory thus deals with selective information problems. In examining the research in semantic information, we observed that Carnap and Bar-Hillel (1952) and Winograd (1972) are perhaps best known for their work in this area.

The aforementioned views of information are two of the three approaches or levels identified in studies of information theory by Weaver. The third level is known as the *behavioral or effectiveness level* and deals with the effect that information has on the person using it. Ackoff (1958) has dealt with information problems at this behavioral level. The work of this project is considered to lie in this area, since we are concerned with the effect of information on *decision making behavior* by a human, in a computer-assisted mode. The Phase 1 report (Llinas et al., 1997) elaborates on each of these viewpoints on informational value.

We also examined work which related informational value to a DM model. Morris (1964) has identified three general requirements of action involved in the decision making process. A decision maker must obtain information about the situation in which he is to act, select among courses of action, and execute this alternative by some specific course of behavior. To effect a meaningful analysis of information, one must examine in detail that which makes decision making such a challenging activity—uncertainty. We concerned ourselves with uncertainty because we will argue that a key role for information is its influence on uncertainty within a decision making process.

The decision maker usually views a complex decision situation in terms of his roles and responsibilities within it, for selecting courses of action (COA), which then lead to possible outcomes. He may be uncertain about what outcomes will occur when a particular course of action is executed. This uncertainty associated with the execution of the alternatives is what Yovits and Abilock call *executorial* uncertainty. A second type of uncertainty identified is *goal uncertainty*. The decision maker may have only a vague notion of the goals to which he aspires, and he may also be uncertain as to the degree to which each of the outcomes will satisfy the various goals. The third type of uncertainty which the decision maker confronts is that concerned with the states of nature. He may not be able to identify all the possible states, but even if he could, he may still be uncertain as to the relationship between the set of states and the other decision elements. This is termed *environmental uncertainty*. A complete model of a complex

⁴ Intelligence analysts similarly concern themselves with "external" information (i.e., that not related to message content) and "internals," which are the content-bearing elements of messages.

decision situation must deal explicitly with all of these types of uncertainty. The conceptual decision model suggested by Yovits and Abilock explicitly recognizes all of the decision elements as well as the associated sources of uncertainty. The following Table 1.1 summarizes these ideas and uncertainty types.

Table 1.1 Uncertainty

Type Of Uncertainty	Aspect Represented
Executional	Outcome Probability, given a selected COA
Goal	Goal Uncertainty (specifically), and/or relationship between Outcomes and Goal Satisfaction
Environmental	State-of-Nature Uncertainty (specifically), and/or relationship between States of Nature and other Decision Elements

1.2.3 Errors In Human Decision Making

In the section above we defined the decision making task in functional terms (i.e., what are the states of nature, the possible actions, and the values of the resulting outcomes). Based on these definitions, we introduced measures of the value of information to the decision system. We then went beyond a functional description to explore the alternatives for allocating the various functions between human and automated decision components, informed by the models of the human functioning in adversarial systems developed in Section 1.1. One aspect of attempting to define these alternatives requires exploring the possible decision making errors, and how they are influenced by humans and automated systems as decision makers.

1.2.3.1 What is Error?

Reason (1990) concentrates on human error rather than error in general, but we can amend his human error definition as follows:

Error is when a planned sequence of activities fails to achieve its intended outcome, when failure cannot be attributed to a chance agency.

Note that this defines three elements:

1. A goal or intention (i.e., the system is purposive or teleological)
2. A set of actions is chosen
3. An outcome of value is implied

These elements can all be seen in the model we elaborated on in Section 5 of (Llinas et al., 1997), where decision was defined as the choice of a series of actions, based on a value structure (intentions) for outcomes. Errors are thus occasions where the "correct" action was *not* chosen. Again speaking specifically of human error, Woods and Roth (1988) state that "error" is a judgment made in hindsight. It is thus assumed possible to evaluate the quality of a decision

(i.e., determine if it was an error) by reference to some external, but lagged, validation criterion where “truth” about the whole situation was eventually discovered. As evidenced by legal inquiries into major system failures (*Challenger*, *Vincennes*, *Bhopal*, *Herald of Free Enterprise*), this external validation is possible in principle, but difficult and costly in practice. This idea is embodied in the concept of a criterion against which decision making performance can be judged. Hollnagel (1998) gives three parts to an error definition:

1. A performance standard or criterion
2. An event or action
3. A degree of volition

He discusses why each of these may be difficult concepts in a theoretical development, but does emerge with a second distinction useful to our thesis: error genotypes and phenotypes. The genotype is a (generic) cause of the error, while the phenotype is the (specific) manifestation of that cause in a particular system. Those who must deal with human error are either trying to infer genotypes from phenotypes (incident investigation) or infer phenotypes from genotypes (incident prediction). In data fusion-supported adversarial systems the immediate need is for incident prediction, so that one must start with the genotypes of erroneous action.

As noted above, errors imply both intention and action. Indeed, Norman’s (1981) early classification of error genotypes divided them into two categories.

Mistakes: Following a wrong intention

Slips: Correct intention but wrong action

Combining this with Rasmussen’s (1987) three levels of human functioning (skill-based, rule-based, knowledge-based), and adding specific memory retrieval failures (lapses), brought Reason (1990) to three basic error types in Table 1.2.

Table 1.2 Error Genotypes, Adapted from Reason (1990)

Level	Error Genotypes
Skill-based	Slips, Lapses
Rule-based	Rule-based mistakes
Knowledge-based	Knowledge-based mistakes

These form the basis of expansions by Reason, Hollnagel and others into more detailed lists or taxonomies of error types.

1.2.4 Cultural Effects On Adversarial Decision Making

No specific works in the area of cultural effects on decision making or especially adversarial decision making were able to be located in our literature search efforts. However, we

found what we believe to be reasonably related works in the management literature having to do with issues in multinational corporations and their various functions and operations. These inputs range from definitions of what culture is to the notions of different societal values and beliefs, recognition factors for decision-makers and some other related topics. Even these works, however, do not address directly the impact of these factors explicitly on decision making. As a first-cut input addressing this topic for our purposes, we simply collated and assembled some inputs from the cited references. We would hope to investigate this subject further in the next phase of work.

1.2.4.1. Basic Notions of Culture: What Culture Is (Hoecklin, 1995):

Culture, as defined by Hoecklin (1995) has the following characteristics.

1. **It Is A Shared System Of Meanings.** Culture dictates what groups of people pay attention to. It guides how the world is perceived, how the self is experienced, and how life itself is organized. Individuals within a group share patterns that enable them to see the same things in the same way and this holds them together. Each person carries within her or himself learned ways of finding meaning in experiences. In order for effective, stable and meaningful interaction to occur, people must have a shared system of meaning. There must be some common ways of understanding events and behavior, and ways of anticipating how other people in your social group are likely to behave. It is only when the meanings do coincide that effective communication can happen.
2. **It Is Relative.** There is no cultural absolute. People in different cultures perceive the world differently and have different ways of doing things, and there is no set standard for considering one group as intrinsically superior or inferior to any other. Each national culture is relative to other cultures' ways of perceiving the world and of doing things.
3. **It Is Learned.** Culture is derived from one's social environment, not from one's genetic make-up.
4. **It Is About Groups.** Culture is a collective phenomenon that is about shared values and meanings.

1.2.5 Trust in Automation

The above sections provide an overview of our previous work, and it can be seen that there are many factors to consider in attempting to develop an understanding of AADM. However, for environments where the human is the "ultimate transducer," (i.e., the means toward enablement of "final" decisions and subsequent action) we hypothesize that there may be a single focusing issue—"trust in automation"—that might be a very high-payoff aspect of the AADM problem. Indeed, we argue, if that final human decision and action depend on the degree to which the human trusts the decision aid output (no matter how developed, corrupted, or displayed), it is that degree of trust which may be the deciding factor in proceeding with a decision and consequent action. Motivated in part by this assertion and in part by the fact that trust is a factor in any case, we focused our efforts on the subject of trust in automation.

1.2.5.1 State of Research:

It will be seen from the sections that follow that the notion of trust is also complex and multi-dimensional, and that rather little work has been done on the particular subject of trust in automation. The starting point turns out to be, as one might expect, the sociological literature on trust among humans. The dimensions and factors affecting human trust do seem to carry over to the case of trust in automation, perhaps because of the human-like metaphor (e.g., *2001: A Space Odyssey's* "HAL") especially ascribed to computers.⁵ Section 2 will elaborate considerably on these notions, examining the human engineering and the sociological literature. Since so few experiments have been done, there is, correspondingly, not much work on the subject of measures and metrics. Section 3 below provides some discussion on this subject. Since our focus is toward human-in-the-loop experimentation, Sections 4 and 5 describe ideas about scenario formulation and implications for an IW laboratory. There are many factors that may influence the formulation of a trusted state. These factors, detailed in Table A.1 in Appendix A, if manipulated by an adversary to possibly influence another actor's state of trust, may lead to information dependencies and vulnerabilities. Additional discussion on these subjects follows in the remaining sections of the report.

1.2.5.2 Behavioral Response to Distrusted Systems:

Our assertion that trust in automation is possibly a deciding issue in determining the final decisions and actions performed by humans implicitly depends on what has been learned or observed to date on how humans cope with distrust in automation. While there has been little work to support or disprove this assertion, that work that has been done shows that there is a significant hysteresis loop that develops when humans suspect malfunctioning decision aids. In the tactical context of IW, this means that an adversary could conceivably take an opponent "off-line" through IW actions that lead to distrust in the opponents decision-aiding system. If this were done at some opportune time, the payoff in a combative sense could be significant. So there is at least limited evidence that the effects of trust can indeed be quite influential in adversarial decision making; we recommend that much more research is needed to both understand and quantify this possibility.

1.3 References

- Ackoff, R. L. (1958). Towards a behavioral theory of communication. *Management Science*, 4(3), 218-234.
- Bar-Hillel, Y. (1955). An examination of information theory. *Philosophy of Science*, 22(2), 86-103.
- Carnap, R., & Bar-Hillel, Y. (1952). *An outline of a theory of semantic information* (Technical Report No. 247). Cambridge, MA: M.I.T., Research Laboratory of Electronics.

⁵ Some argue that human-like behavior of computers clearly follows from the fact that they are programmed by humans, and so this relationship is more than a metaphor; it is real (but this is a separate subject).

- Clarke, A. C. (1968). *2001—A space odyssey*. New York: New American Library.
- Denning, P. J. (Ed.). (1990). *Computers under attack: intruders, worms, and viruses*. Reading, MA: Addison-Wesley Publishing Co.
- Dept. of the Army. (1994). *FM 100-6, Information operations*. Washington, DC: Headquarters, Dept. of the Army.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, Special Issue on Situation Awareness*, 37(1), 32-64.
- Hoecklin, L. (1995). *Managing cultural differences*. Wokingham, UK: Addison-Wesley.
- Hollnagel, E. (1998). *CREAM—Cognitive reliability and error analysis method*. New York: Elsevier Science.
- Kaempf, G. L., Klein, G. A., Thordsen, M. L., & Wolf, S. (1996). Decision making in complex naval command-and-control environments. *Human Factors*, 38(2), 220-231.
- Libicki, M. C. (1997). *Defending cyberspace*. Washington, DC.: National Defense University, Institute for National Strategic Studies.
- Llinas, J., Drury, C., Bialas, W. & Chen, A. (1997). *Studies and analyses of vulnerabilities in aided adversarial decision-making: Final Phase 1 Report*. Buffalo, NY: SUNY at Buffalo, Center for Multisource Information Fusion.
- Luoma, W. (1994). *Netwar: The other side of information warfare*. Newport, RI: Naval War College. (DTIC Report No. ADA 279 585)
- MacKay, D. M. (1952). In search of basic symbols; The nomenclature of information theory. In Heinz von Foerster (Ed.), *Cybernetics: Transactions of the Eighth Congress* (pp. 181-221; 222-235). New York: Macy Foundation.
- Morris, C. (1964). *Signification and significance*. Cambridge, MA: M.I.T. Press.
- Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, 88(1), 1-15.
- Rasmussen, J. (1987). Reasons, causes and human error. In J. Rasmussen, K. Duncan and J. Leplat (Eds.), *New technology and human error* (pp. 193-301). New York: John Wiley & Sons.

- Reason, J. (1990). *Human error*. Cambridge, UK: Cambridge University Press.
- Riley, V. (1989). A general model of mixed-initiative human-machine systems. *Proceedings of the 33rd Human Factors Society Annual Meeting, Vol. 1*, 124-128.
- Shannon, C. & Weaver, W. (1964). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.
- Sheridan, B. (1976). Toward a model of supervisory control. In T.B. Sheridan and G. Johanssen, (Eds.), *Monitoring behavioral and supervisory control* (pp. 271-281). New York: Plenum Press.
- Stein, G. J. (1995, Spring). Information warfare, *Airpower Journal*, 31-39.
- Szafranski, R. (1995, Spring). A theory of information warfare. *Airpower Journal*, 56-65.
- Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, 3(1), 1-172.
- Woods, D. D. & Roth, E. M. (1988). Cognitive systems engineering. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 3-43). Amsterdam: Elsevier.
- Yovits, M. C. & Abilock J. (1974). A semiotic framework for information science leading to the development of a quantitative measure of information. *Information Utilities: Proceedings of the 37th American Society for Information Sciences (ASIS) Meeting, Vol.11*, 163-168.

2.0 CONCEPTS AND DEFINITIONS OF TRUST AND RELATED NOTIONS

As the degree of complexity and intercommunication of society increases, the degree of dependence on others will become greater. Almost all of everyday decisions involve trusting someone else, or sometimes, depending on others completely. It is probably no exaggeration to say that our society is becoming more and more based on trust. The study of trust has a long history; however, that history is not rich in empirical studies. Deutsch (1958, 1960) made the first traceable attempt to define and examine potential characteristics of the term forty years ago. Only a few studies since then have attempted to understand the role of trust in interpersonal relationships. Recent studies in a sociological context (e.g., Barber, 1983; Rempel, Holmes & Zanna, 1985; Holmes, 1991) define trust to be a multi-factorial concept which has a direct implication for human trust in automated systems. Based on the definitions from a sociological perspective, two essential studies (Muir, 1994; Lee & Moray, 1992) showed the role of human trust in a continuous process control environment, producing results generally consistent with those from the sociological study of trust, despite some exceptions. However, none of this research has examined the possibility of applying the concept of human trust in the IW domain even though the domain itself has drawn significant research attention. Therefore, to bridge the gap between the sociological aspects and the IW environment, research on trust in human-machine interaction will be reviewed.

2.1 Overview of the Sociological Literature

Studies which investigate the role of trust in human relations suggest that trust is a *sine qua non* (one of the foundations) and an ultimate element of such relationships. A typical example, which shows its importance in human relationships, is a study using the dyadic (pairwise) or interpersonal trust scale by Larzelere and Huston (1980). They measured the mean dyadic trust scores between couples and showed that such scores depended on the couples' relationship development status. Specifically, trust increased as the relationship developed from casual dating couples through newlyweds to longer married couples, even though the sample sizes of each relationship status were different. As might be expected, separated or divorced couples showed the lowest trust scores, newlyweds, the highest, with the scores of the longer married couples very close to those of the newlyweds.

Along with documenting the importance of trust in human relations, a few attempts have been made to define and characterize the concept of trust in interpersonal relationships. Deutsch (1958) made the first notable attempt and characterized trust as involving two notions: expectation (predictability) and motivational relevance. Before discussing Deutsch's work, we review some definitions.

Webster's Third New International Dictionary (1993) defines trust in four ways:

1. assured reliance on a person or thing
2. dependence on something future or contingent

3. an equitable right or interest
4. a charge or a duty imposed in faith or confidence or as a condition of some relationship
 - 4.1 something committed or entrusted to one to be used or cared for in the interest of another

In Deutsch's definition, two factors are recognized as the basis for trust. First, as in Webster's second definition, the person to be trusted should be predictable. However, predictability is not sufficient to capture all aspects of trust. Sometimes, one must predict an event (or someone's future behavior) without being able to rely on historical information. Secondly, without the interest of the partner who gives trust, even consistent behavior becomes meaningless, which is the same as Webster's fourth (4.1) category. Therefore, a human should be fully motivated to pay attention to the partner's past behavior.

These characteristics are reflected later in Rotter's (1967) definition of trust as "an expectancy held by an individual or group that the word, promise, verbal, or written statement of another individual or group can be relied on." In his next study, Rotter (1971) subdivides "expectancy" by embracing a social learning theory to include situations where humans are not familiar with a situation and are forced to generalize and deduce from past experience. He concludes that expectancy is a function of two distinctive types of experience: a specific experience and a generalized expectancy resulting from the generalization from related experience. While specific expectancy is defined as a function of degree of experience of a specific situation, it is understood that the degree of novelty, ambiguity, or unstructuredness of a particular situation can affect human trust. At this point, however, it has not yet been proven either how much the similarity or the degree of association between the current situation and the expectancy stored in the human operator's mental model can affect human trust level or whether its impact on human trust is positive or negative. This attribute can be labeled "familiarity," one of the seven attributes suggested by Sheridan (1980) as major characteristics of trust, as shown later. Other definitions of trust in human relationships are as follows:

Scanzoni (1979): An actor's willingness to arrange and repose his or her actions on another actor because of confidence that other will provide expected gratification.

Larzelere & Huston (1980): A belief by a person in the integrity of another's behavior.

Barber (1983): The expectation of the persistence and fulfillment of the natural and the moral social orders, expectation of technically competent role performance, expectation that partners in interaction will carry out their aforementioned characteristics (persistence, technically competent performance, and fiduciary responsibility).

Rempel, Holmes, & Zanna (1985): A generalized expectation related to the subjective probability an individual assigns to the occurrence of some set of future events.

Rempel and Holmes (1986): The degree of confidence one feels when one thinks about a relationship.

Two of these definitions contain critical aspects of trust which can be used to examine human trust in automation from a human factors perspective. Barber (1983) described three types of expectations related to the three dimensions of trust: persistence of natural and moral laws, technically competent performance, and fiduciary responsibility. According to Barber, persistence of natural and moral laws provides a foundation of trust by establishing a constancy in the fundamental moral and natural laws. Persistence of natural and moral laws reflects the belief that "... the heavens will not fall," and that "... my fellow man is good, kind, and decent" (Barber, 1983; p. 9). These expectations provide the basic conditions for social and physical interactions.

Technically competent performance, on the other hand, supports expectations of future performance based on capabilities, knowledge, and expertise. This dimension of trust refers to the ability of the other partner to produce consistent and desirable performance and can be subdivided to include three types of expertise:

- Everyday routine performance
- Technical facility
- Expert knowledge

These types will be explained later in conjunction with another human factors error nomenclature.

Barber's third dimension of trust, fiduciary responsibility, concerns the expectation that people have moral and social obligations to hold the interests of others above their own. Fiduciary responsibility extends the idea of trust beyond that based on performance to one based on moral obligations and intentions. This dimension becomes important when agents cannot be evaluated because their expertise is not understood, or in unforeseen situations where performance cannot be predicted. Here expectations depend upon an assessment of the intentions and motivations of the partner, rather than on past performance or perceived capabilities. Fiduciary responsibility also implies that the actors in a relationship exist in a *cooperative* framework. In the cases of interest here, the human-automation relationship, as embodied in the decision-aiding software built by the "friendly" force agents, would also be called cooperative. However, importantly, we are also concerned with hostile information attacks on this supposedly friendly software, which clouds the overall nature of the human-computer relationship.

In addition to the dimensions of trust proposed by Barber (1983), Rempel, Holmes, and Zanna (1985) emphasized not only components of interpersonal trust, but also the dynamic characteristics of trust toward a partner, regarding trust as a generalized expectation related to the subjective probability an individual assigns to the occurrence of some set of future events (Rempel, et al., 1985). The three major components of Rempel et al.'s (1985) definition of trust are predictability, dependability, and faith. According to Rempel et al., predictability, which represents the consistency of recurrent behavior and the stability of the social environment, forms the basis of trust *early* in the relationship. As interpersonal relationships progress with further experience, dependability, which represents a more common understanding of the stable dispositions, becomes an important basis of trust between humans and focuses on an evaluation

of the qualities and characteristics attributed to the partner. In this stage, therefore, the centroid swings and shifts away from the evaluation of the partner's dispositional attributes to the partner, *per se*. Faith, on the other hand, describes the aspects of trust or belief that must go beyond the available evidence to permit the truster to accept a given supposition as truth. Even though this idea still lacks either the motivational or the degree of self-confidence factors of the person who is giving trust, which were considered in a later study, it embraces the basic important notions of trust.

2.2 Overview of the Human Factors Engineering Literature

So far, it has been shown that trust is a multi-factorial concept, whether using the dictionary, speculation (Barber, 1983; Rempel et al., 1985; Zuboff, 1988), or our own expectation. To apply the studies of trust in a sociological context to the human-machine environment, we may need some transformation or different interpretation from the human-human environment. While many studies (e.g., Larzelere & Huston, 1980) have justified the use of measurements using rating scales, it is also obvious that different measurement schemes, (e.g., percentage of time that operators used in automatic controllers, and frequency of monitoring activities) should be developed to predict an operators' actions in a human-machine interaction environment. There have been only a few studies of trust from this perspective. In this section, these studies will be summarized to establish the analogy from trust in automation to trust within the IW environment.

2.2.1 Supervisory Control and Automation

Automation has played an important role in supporting human and system performance in complex modern systems, such as in those found in aviation and process control settings. Automation technologies have relieved the burden on the human operator to perform under difficult or dangerous physical conditions and have also augmented human abilities to gather information, such as providing a non-destructive method for aircraft inspection (McMaster, McIntire & Mester, 1986) and autopilot systems (McClellan, 1994). The existence of benefits from the introduction of automation technology is undeniable. In short, automation technology provides us a convenient and economical way of living.

However, this is not the case in industrial applications, where a sophisticated environment generates requirements for broad human operator activities to support or control the automated processes. The advent of automation has changed the role of the human operator from that of performing direct manual control to that of managing different levels of computer control. Functions are often automated within the current available technology and economical constraints without considering or defining the human operator's role as a subsystem (e.g., Bainbridge, 1983). Thus, human operators are left to assume the role of a supervisory controller, interacting with the system through different levels of manual and automatic control (Sheridan & Johanssen, 1976). Therefore, the human operator must understand how to interact with system computers, how the computers work, how to respond based upon the output from computers, and how and when to intervene in the process if the process fails. Many suggestions for distributing and allocating a variety of system functions across the two essential subsystems, automation and human operators (Levis, Moray & Hu, 1994), have been made to improve the overall

performance of human supervisory controllers (see Sheridan, 1992, 1997, for details). These range from ensuring stimulus-response compatibility to developing an internal model of the human operator (Kantowitz & Campbell, 1996).

Sheridan (1980), and Sheridan, Vamos, and Aida (1983) emphasize the importance of human trust in automation as playing a key role in determining the level of a human operator's reliance on and the degree of intervention in automation and appropriate use of automation (see also Parasuraman & Riley, 1997). Although those research projects concentrated on human trust in automation based on the general understanding of supervisory control tasks, the importance of the trust concept is applicable and has been applied to other domains, such as computer supported cooperative work (Jones & Marsh, 1997; Christianson & Harbison, 1997), decision making in management (Lerch & Prietula, 1989), medical diagnosis expert systems (Moffa & Stokes, 1996) and also computer security problems (Beth, Borcharding & Klein, 1994).

2.2.2 Models of Trust

Sheridan (1980) introduced the idea of the importance of human mystification with (and misplaced trust in) automation as one of the seven factors in the alienation of people from technology. Based on the sociological definitions of trust described in Section 2.1, Muir (1994) constructed a model of human trust in automation by incorporating the dimensions of trust proposed by Barber (persistence of natural laws, competent performance, and fiduciary responsibility) and three more dimensions of trust (predictability, dependability, and faith) from Rempel, Holmes, and Zanna (1985). According to Muir's interpretation, all three of Barber's (1983) meanings of trust seem applicable to the human-machine relationship and become a basis for the framework. Thus, Muir's work, if taken as a definitive reference, forms the first "transition" of the concepts of trust in human-human relationships to human-automation relationships. Muir's work, in fact, suggests that this is an identity-transform, (i.e., that the human-human trust concepts extend directly to the human-automation relationship case). Muir (1994) also identified one of Barber's aspects of trust, technical competent performance, with Rasmussen's (1983) taxonomy of behavior: skill-based, rule-based, and knowledge-based behavior; this relationship is shown in Table 2.1.

Table 2.1 Association of Barber's Technical Competent Performance to Rasmussen's Taxonomy

Barber's technical competent performance	Rasmussen's taxonomy
Everyday routine performance	Skill-based
Technical facility	Rule-based
Expert Knowledge	Knowledge-based

Interpreting Rempel, Holmes, and Zanna's (1985) model as a hierarchical stage model, able to account for changes in operators' trust as a result of experience on a system, where trust develops over time, Muir produced a framework by crossing Barber's (1983) dimensions of trust, with Rempel et al.'s (1985) framework; this crossed relationship among these various dimensions is shown in Table 2.2.

Table 2.2 Muir's Framework for Studying Trust in Supervisory Control Environments, Produced by Crossing Barber's (1983) Taxonomy of Trust (Rows) with Rempel, Holmes, and Zanna's (1985) Taxonomy of the Development of Trust (Columns). Adapted from Muir (1989).

Basis of Expectation at Different Levels of Expertise			
Dimensions from Barber(1983)	Dynamic dimensions from Rempel et al. (1985)		
Expectation	Predictability (of acts)	Dependability (of dispositions)	<i>Increasing time</i> Faith (in motives)
Persistence			
Natural physical	Events conform to natural laws	Natural is lawful	Natural laws are constant
Natural biological	Human life has survived	Human survival is lawful	Human life will survive
Moral social	Humans and computers act "decently"	Human and computers are "good" and "decent" by nature	Humans and computers will continue to be "good" and "decent" in the future
Technical competence	j's behavior is predictable	j has a dependable nature	j will continue to be dependable in the future
Fiduciary responsibility	j's behavior is consistently responsible	j has a responsible nature	j will continue to be responsible in the future

In Rempel et al.'s (1985) framework, predictability is the factor dominating *early* in a relationship, dependability dominating later, and faith dominating in a mature interpersonal relationship. Muir suggested a hypothetical trust model for complex systems:

$$T_i = E_i(P_j) + E_i(TCP_j) + E_i(FR_j) \\ = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_1X_2 + B_5X_1X_3 + B_6X_2X_3 + B_7X_1X_2X_3$$

Where i = individual holding the expectation to recognize explicitly that trust,

j = referents (complex system),

B_{0-7} = parameters,

X_1 = P (persistence),

X_2 = TCP (technically competent performance), and

X_3 = FR (fiduciary responsibility).

That is, trust (T) is the expectation (E) held by a person (human operator) of a system (i) of the persistence (P) of the natural and moral orders, and of technically competent performance (TCP), and fiduciary responsibility (FR) from a partner, referent, (j) of the system and is related to objective measures of these quantities and the various interaction effects.

In addition to providing a broad theoretical framework for studying trust, Muir and Moray (1996) also conducted two experiments which contribute to an understanding of trust between humans and machines, using a simulated pasteurization plant which was controlled by either manual or automated controllers. The process, displayed on a visual display terminal in a mimic diagram, showed the overall flow of milk pasteurization to assist the operator's understanding of the process. There were two important automatic subsystems for controlling the process in these experiments: the pump subsystem, and the heating subsystem. The former controlled the level/amount of raw material flowing into the process to be pasteurized, while the latter defined the amount of heat required to pasteurize the milk. These subsystems could be controlled either

by manually typing the target pump rate/amount of heat applied to the milk or by engaging an automatic mode. The automatic mode reset the pump rate to match the flow rate of milk into the system and reset the heat transfer rate to keep the temperature of milk within the predefined range. The information available to the human operators through the mimic diagram of the entire pasteurization plant was in digital form and indicated the level of pumping and heating rate for the pumping and the heating subsystems, respectively. The display did not provide any decision alternatives. After experience in controlling the plant with whichever strategy they chose, operators were asked to rate their trust in automation, *in general*.

Two fundamental hypotheses were made for the experiments:

1. As the level of human operator's trust increases the operators would engage in the automatic mode more, measured by the percentage of the time engaged in the automatic mode.
2. Also, as the level of human operator's trust increases, the operators would monitor the automation less.

The first experiment failed to show a strong relationship between trust and the percentage of time that operators used the automatic controller, but still demonstrated the operators' ability to generate subjective ratings of trust. The reason why it failed was that the human operators engaged in extensive manual control to maximize the output performance. In other words, they showed a ceiling effect in which the human operators committed to control the automated systems in manual mode only (almost 100% of time). This was probably due to the reward structure based on the performance, in which the human operator who performed the task with the most output was promised to be rewarded. With the consistent preference for manual control (ceiling effect), the assessment of the relationship between operator's trust in and the percentage of the time that operators used the automatic controller became meaningless.

In the second experiment, one of the automated subsystems, the heating subsystem, was changed to allow only automatic mode operation while the other subsystem, the pump subsystem, was controllable in either mode. Thus, the human operators were not allowed to intervene with the heating subsystem. However, the automated heating system was so highly reliable that the human operator's task was to monitor the automated system.

The second experiment showed a strong positive relationship between trust in the automatic controller and its use of the feedstock pump, and an inverse relation between trust and monitoring. That is, as the level of the human operator's trust increased, they used the automatic mode more than manual mode and were less occupied with automation monitoring activities. More importantly, it seems that the reason why they failed in the first experiment and were successful in the second was because of the increased specificity of the trust rating. As opposed to generating an overall trust measure in the feedstock pump, operators were instructed to generate their trust in a specific object, i.e. automatic controller of the feedstock pump. Two conclusions from these experiments are that operators are able to generate subjective ratings of their trust in automation and that trust ratings are correlated with specific characteristics of automation which were defined by Lee (1992).

Rempel et al. (1985) describe their model as a hierarchical stage model, "but only in the sense that we suspect that there is a developmental progression in terms of the time and emotional investment required to establish each component and in terms of level of attributional abstraction each demands" (p. 98). They found faith to be the most important aspect of trust in close interpersonal relationships. Applying Rempel et al.'s (1985) model of trust between humans, Muir and Moray (1996) found results consistent with Rempel et al. (1985), in that three characteristics (predictability, dependability, and faith) were fundamental attributes of human trust and were developed over time. Muir and Moray's (1996) result, however, also showed that faith was a better predictor of trust, *early in a relationship, but not late*. Recall the finding from Rempel et al. (1985) that faith is the important factor *late* in a relationship. This finding seems to represent a difference between human-machine relationship and the human-human relationship, even though Rempel et al.'s dimensions of trust in interpersonal relationships has direct implications for human-machine settings. As Lee (1992) pointed out, this must be because the initial instructions regarding a machine provided to human operators is "its intended use." In human relationships, on the other hand, it may take years of experience to understand a human partner's intention and to develop faith in the relationship. It might suggest that the human operator's trust toward automated systems may tend to develop very rapidly toward faith, or, perhaps that a human operator may start at a higher level of trust when dealing with automated systems than when dealing with people. Faith, then, should be redefined in operational terms rather than interpersonally. Thus, faith in the human-computer environment seems to be compatible with the basic, generalized, and somewhat ambiguous trust in automated systems, which is one of the expectancies classified by Barber (1983).

Following the groundwork laid by Muir, Lee (1992) extended Muir's work to investigate various effects of system failures. He used Muir's model of trust and proposed dimensions of trust and the relationships between the different dimensions of trust.

Table 2.3 shows a comparison between Lee's and others' dimensions of trust. According to Lee, four dimensions of trust are defined and matched with other sociological definitions of trust, rather than counter-balanced against each other in an orthogonal manner as Muir did with Barber's and Rempel et al.'s dimensions (shown here in Table 2.2). The first dimension is the foundation of trust, representing the essential assumptions of natural and social order that makes it a cornerstone for other dimensions of trust. Also, this dimension of trust corresponds exactly to the persistence of natural laws described by Barber (1983). The second dimension of trust, performance, describes the expectation of consistent, stable, and desirable performance or behavior. The third dimension, process, is characterized as depending on an understanding of the underlying qualities or characteristics that govern behavior, such as dispositions or character traits. In human-machine interaction environments, this would be a control algorithm or a data reduction method that controls how the system behaves. The final dimension of trust, purpose, rests on motives or intents. However, a machine's motives or intents are the reflections of designers' intentions or purposes in creating the system. This classification of the dimensions of trust seems to lead to the conclusion that machines or automated systems are easier to trust. Before making a hasty conclusion, we may have to consider Rotter's dimensions of trust which derive from specific and generalized notions of expectancy. Regardless of the specific expectancy, we, as human beings living in the '90's, have experienced all variety of automation systems, ranging from a trivial system such as automobile's cruise control system to a very

complex system. Therefore, we may tend to regard automated systems with a certain amount of trust in advance, before even dealing with the system. These opposing viewpoints will be discussed later along with Muir and Moray's (1996) results.

Table 2.3 Proposed Dimensions and Relationship Between the Different Dimensions of Trust. Adapted from Lee (1992)

Lee (1992)	Barber (1983)	Rempel, Holmes, and Zanna (1985)	Zuboff (1988)
Foundation	Persistence of Natural Laws	Not applicable	Not applicable
Performance (consistent, stable, etc.)	Technically Competent Performance	Predictability	Trial-and-Error Experience
Process (understanding behavior)	Not applicable	Dependability	Understanding
Purpose (understanding intent)	Fiduciary Responsibility	Faith	Leap of Faith

To investigate the dynamic characteristics of trust, Lee (1992) conducted an experiment with a structure very similar to the one used in Muir's experiment. In addition to the several characteristics which were investigated in Muir's, such as the effects of magnitude of error, effects of types of error (constant vs. variable), two different types of error were considered here; transient, and chronic faults. Transient faults are considered as sudden irrational behaviors of the machines. In these experiments, transient faults were deliberately programmed to appear only once. Chronic faults, on the other hand, are regarded as eventual mechanical, automation failures and happened throughout a trial. When the faults occurred, the actual automated system failed to reach its requested rate whether controlled by the operator or the automatic controller. Fault magnitudes were predefined at four levels (15%, 20%, 30%, and 35%). The operators' levels of trust were measured in response to a questionnaire, using a subjective rating scale such as was used by Muir (1989). These questions were intended to ask how the operators felt about the system's characteristics (predictability, dependability, and faith), as defined by Rempel et al. (1985). These questions were (Lee & Moray, 1992):

1. To what extent can the system's behavior be predicted from moment to moment?
2. To what extent can you count on the system to do its job?
3. What degree of faith do you have the system will be able to cope with all system's states in the future?
4. Overall, how much do you trust the system?

As shown, these are very direct questions to the operators about their feelings concerning the overall system. Although Lee (1992) himself realized that the low degree of specificity about the system being rated for trust was the reason why Muir (1989) failed to show the relationship between the human operator's trust and the automation usage, he used very generalized questions.

Lee (1992) found a mathematical model of trust using an autoregressive moving average vector form, as follows:

$$\text{Trust}(t) = \phi_1 \text{Trust}(t-1) + A_1 \text{Performance}(t) + A_1 \phi_2 \text{Performance}(t-1) + A_2 \text{Fault}(t) + A_2 \phi_3 \text{Fault}(t-1) + a(t)$$

Where

t : time subscript

A₁ : The weighting of system performance

A₂ : The weighting of the occurrence of a fault

φ_i : Autoregressive moving average vector form time constraints

a : random noise perturbation

He found out, first, that operators were able to maintain their high level of overall performance, despite the effects of a fault in the automatic controller. As might be expected, there was a loss of trust in automation resulting from faults in the automatic controller, and the recovery of trust was slower than recovery in performance. Lee called this "inertia" (we called this a "hysteresis loop" in the last report). This result is consistent with Lerch and Prietula (1989), who examined the effects of attributional qualities of a source (i.e., the pedigree) on human decision making in traditional financial management decision problems. They also found that it was more difficult to recover trust after a failure, given as wrong advice, than to build trust initially. The level of performance measured by the level of confidence in the decision the subjects made deteriorated after the wrong advice, and never returned to the level of performance where it was before the wrong advice, even at the end of trials. Second, the magnitude of the automatic controller's error had no differential effect on overall performance but the change in trust was positively related to the magnitude of the error. Third, a decrease in trust of the automatic controller caused no decrease in its use. Lee and Moray (1994) hypothesize this may have been a function of the level of the operator's self-confidence. Trust, paired with self confidence, may provide a better explanation for operator's use of a decision aid, after it fails, than trust alone. The use of an intelligent system seems to depend on the human's perception of their own capabilities, as well as their perception of the systems performance. Thus, combining these qualities seems to lead to better joint performance.

These results have potentially significant implications for the IW case. If an adversary can cause loss of trust (e.g., by perturbing system information), he can effectively remove the human from system control for some meaningful length of time.

2.3 Characteristics of Trust

In the previous section, we have reviewed some definitions and models of trust from both sociological and human factors engineering perspectives. These definitions and models of trust contain many aspects, suggesting trust is a multi-factorial concept. Now, we must go beyond the suggested description of trust models that is usually applied to both human-human relationships, and human-machine interaction, to eventually apply trust concept into IW domain. Specifically, we need to explore the possible characteristics of trust that may contribute to change in human behavior.

2.3.1 Characteristics of the "Other Actor"

Sheridan (1988) addressed a number of meanings of the term trust, examining how trust affects the operator's use or nonuse of automation features when the occasion arises, and suggested seven attributes of trust, or as he pointed out "perhaps these are better stated as causes of trust," in command and control systems. These are described below:

1. *Reliability*: This implies a system of reliable, predictable, and consistent functioning. Almost all definitions addressed in the previous section placed this attribute as the first step toward development of trust. A person who behaves in a consistent manner will be trusted easily. Further, Sheridan (1988) emphasized this attribute as "conditioned to trust" under which those events which happened before in particular circumstances will occur again in the future. Therefore, the attribute has the same meanings as Rotter's (1971) term "specific expectancy." As Rotter (1971) suggested, people can generalize their expectancy from similar past experiences. Muir and Moray's (1996) experiments supported the importance of reliability affecting the level of trust in automation, and stated that trust is affected by the error magnitude. The characteristics addressed by other researchers which have a similar connotation are predictability, confidence, persistence, trustworthy, and expectancy.
2. *Robustness*: Robustness supports expectations of future performance based on capabilities and knowledge not strictly associated with specific circumstances that have occurred before. Robustness can be stated as "meaning demonstrated or promised ability to perform under a variety of circumstances" (Sheridan, 1988). It is the same as Barber's and Muir's concepts of generalized competence
3. *Familiarity*: Often a person confronts a situation or an object with a high degree of novelty, but still feels familiar with and sometimes comfortable to deal with, the situation. Often caused by either a naturalistic or inherent cultural expectation, familiarity may not cause any exploratory risk-taking behavior to diagnose the situations, or to identify objects whether new or familiar. Consequently, it may induce biased decision-making. However, the fact that familiarity is not based on any scientific knowledge or expertise and tends to be inherited from those who have cultural similarity with us, the person who is confronting an unfamiliar or unanticipated situation / object will be very vulnerable to deception. Unlike other industrial settings where unanticipated, and so unfamiliar, events are sometimes confronted by human operators, human operators in military command, control, communication and information system (C³I) may not have been exposed and trained to any unanticipated events. The following quote from Sheridan (1992) gives us a good understanding of how naïve the human operators in the C³I systems can be for any unfamiliar events.

"...The author has observed some very limited military "war games" and has noted an interesting tendency to avoid the unexpected and to have the "bad guys" behave in rather stereotypical ways, the rationale being that the commanders cannot learn proper procedure and doctrine if events are too chaotic." (p. 351)

Familiarity will be discussed later, in more detail, from the perspective of the ecological interface design in conjunction with "Explication of Intention."

4. *Understandability*: Understandability is neither totally the same as nor completely different from familiarity. Familiarity does not guarantee understandability, or vice versa. The

construct of understandability is equivalent to developing an appropriate mental model, possibly with the aid of familiarity. In designing a machine to aid a human operator, understandability usually is affected by the degree of transparency of the system in which the operator can "see" through the interface to the underlying system. Opaque machines or interface media will prevent the operator not only from trusting the machines, but also from engaging in problem-solving activities in cases of warnings or mishaps. Thus, any means through which an adversary could corrupt the graphical user interface or other user interface functions, so as to confound the user's ability to understand a system, would lead to distrust. For example, corruption of "Help" files could cause such distrust.

5. *Explication of Intention*: Instead of leaving a person in a position where the covert meanings have to be discovered and understood from the system's behavior, this attribute allows people to trust others over those who just perform tasks. However, current technological improvements in the design of intelligent computers are not yet well enough developed to allow human operators to communicate using higher level intentions. Unless we develop intelligent machines which can specify their intentions for future actions outright, we will have to rely on currently available technologies (e.g., in the form of symbols, short statements, or a combination of both which are pre-programmed by system designers and often are not well suited for transferring their intentions to the human operators). Therefore, we are often forced to trust (or not to trust) based on a symbolic medium through which one produces effects and on the basis of which one derives an interpretation of "what is happening." Zuboff (1988) examined this conflict of using symbolic media instead of oral presentation. This characteristic will be discussed later in more detail in conjunction with "Familiarity" from the perspective of the ecological interface design.
6. *Usefulness*: In a sociological context, this attribute is the same as the notion of motivational relevance, which is also the same as Webster's third category—an *equitable right or interest*. From a human factors perspective, on the other hand, usefulness of data or machines means responding in a useful way to create something of value for operators, eventually developing into trust. In fact, one branch of decision theory, "utility theory," is explicitly based on such values. This, however, raises a question: does usefulness of data ensure the quality of decision making, or make human operators dependent on the decision aids, eventually trusting them? In other words, does data value help decision performance, induce trust, or both? Tversky and Kahneman (1974) argued that people tend to behave heuristically rather than using the estimated utility. Humans also tend to express overconfidence for positive-outcome events and the results revealed a curvilinear relationship between base rates (in probability) of possible outcomes and overconfidence (Pulford & Colman, 1996). Klein (1997) also argued that humans tend to use recognition-primed decision making processes rather than evaluating the utility of all possible outcomes. The RPD model is an example of a naturalistic decision making model which is claimed to be the way people use their experience to make decisions in the context of a task. In his argument, Klein (1997) claimed that the function of the RPD model is to describe how people can utilize their experience to make good decisions without comparing the strengths and drawbacks of alternative choices of action, and provides one of the few approaches to understanding human decision making which does not focus on evaluating trade-offs among discrete possible choices or actions. Thus, the model does not reflect any metacognitive processes (i.e., attentional resources and

memory) which are important components of utility theory. This does not deny the usefulness of utility theory as a concept but rather questions how such a concept is used in practice. These effects may also be coupled to the relevance of the data or information. Humans will clearly disregard the presence of data they consider irrelevant, and if this is the context in which heuristic behaviors occur, it is quite understandable. But if not, then the question of the value of information in decision making becomes less certain, and this has important implications in the IW case.

7. *Dependence*: Sheridan (1988) argued that trust should precede dependence rather than follow it. Despite this seemingly rational deduction, we can find many situations in which humans totally depend on the systems even without developing trust (i.e., people trust automation in such situations because they are not confident in their own skills and see no clear alternative to trusting the system). It seems clear that this attribute plays a major role in developing trust and in determining the degree of human intervention in the system. The causal relationship between dependence and operator's trust, however, is still ambiguous.

2.3.2 Characteristics of Data Communicated

We have been discussing the major factors determining the level of human intervention in a human-automation interaction environment. To intervene within the system, human operators must have knowledge, skills and awareness of a surrounding environment that changes dynamically. Increased awareness of the surrounding environment is accomplished by watching or monitoring automation through the displays or interfaces provided, which is the only way for human operators to interact with the system. Before our society acknowledged the importance of information as a medium that can lose its value or integrity, the displays were commonly regarded as simple devices representing the behavior of the machines with which human operators were concerned. Realizing the potential for information degradation or corruption by other individuals or parties, however, raises several questions from a human factors perspective.

1. How can we detect abnormalities of degraded or corrupted information?
2. How much do we trust the displays, as well as automation?
3. What kind of data do we inherently trust?

When we measure a human operator's trust in automation, we already hypothesize in advance that the level of trust in automation is the same as the level of trust in the information presented. This provides us with a single methodology to measure a human operator's trust in automation without questioning the human operator's ability to separate the data from its source. However, this is not the case in the IW environment. The whole C³I system is by nature subject to malevolent interference by a powerful adversary. Data can be corrupted or degraded by adversaries, and consequently, may induce decision making errors.

These questions imply that we should define the role of displays as independent agents rather than as part of the entire system. Displays can create or produce faults either through manipulation by other individuals or through the mechanical malfunction of the displays themselves. Also, remote sensors collect and fuse data from objects in the real world with which human operators actually deal. Will knowing the precise level of the human operator's trust in

the objects or the automation allow us to assume that the level of their trust in the data carries the same value? Suppose the level of human operator's trust in automation is 90%. Does that apply to the displays, too? In other words, we are questioning the human operator's ability to separate his/her level of trust in displayed data from his/her level of trust in the underlying automation (typically, software processes).

We have found only one study (Muir & Moray, 1996) which investigated the effects of display faults in conjunction with automation faults (Muir labeled these "control faults"). Three types of control failures ("exact," "constant fault added," "variable fault added") were manipulated, in conjunction with three types of display failures ("honest," "constant fault added," "variable fault added"). While the "exact" control property would accomplish the requested target rate feed either from automatic or manual mode, the "honest" display property showed the information to human operators, regardless of its quality. Thus, it presented the information even if it contained faults caused by "control" properties (constant fault added, variable fault added). Figure 2.1 represents the way in which human operators interacted with automation in the experiment.

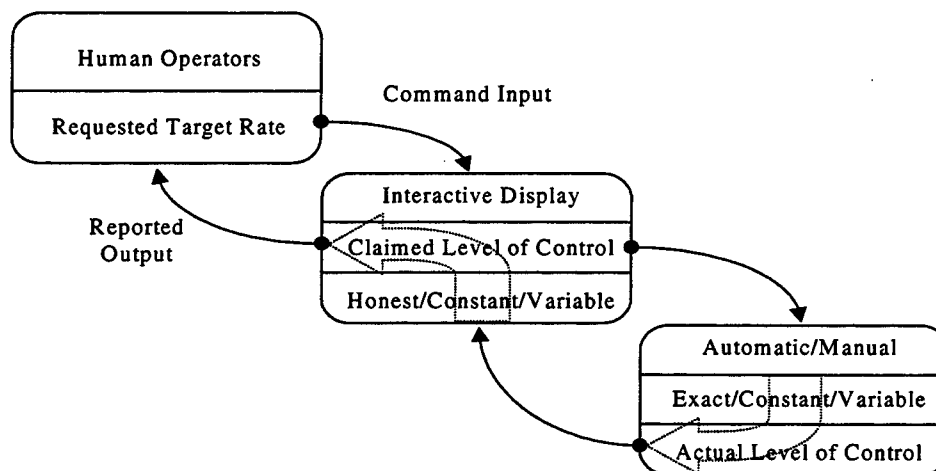


Figure 2.1 Conceptual diagram representing the interactive process display interface between the human operators and the automated system used in Muir and Moray (1996).

Muir concludes that a human operator's trust in displays was affected by both the control and the display manipulation, but that the effects were not additive. The result shows that the honest display was trusted over the two displays that had the constant or variable faults added, across the various types of control properties. The two other displays were trusted nearly equally, although the constant fault display was trusted slightly more than the one with a variable fault added. Common sense seems to be enough to understand this phenomenon, in which bearing constant fault presumably shows more consistent behavior than displaying variable faults. Parasuraman, Molloy, Mouloua and Hilburn (1996), however, produced results contradictory to Muir's results. Parasuraman et al. (1996) investigated the effects of automation reliability and consistency on the human operator's detection rate of automatic failure. They discovered that the monitoring performance of the variable-reliability group was significantly higher than that of the constant-reliability group. Of course, the absolute level of automation reliability also affects

monitoring performance. We may infer from these two results that automated systems with constant reliability may associate with low monitoring performance and then overtrust the automated system. This, then, is a typical example of miscalibration of trust—complacency.

Here, the result from Muir and Moray (1996) shows an interesting potential characteristic in which human operators rate their trust in two automated subsystems. The human operators calibrated their trust very well according to the different types of faults introduced. In the condition in which both agents (control and display) behaved perfectly (“exact” and “honest,” respectively), the level of the human operator’s trust was rated nearly perfect (100%). Among the nine experimental conditions tested (3 levels of control fault properties by 3 levels of display fault properties), two conditions appeared to be exactly the same to human operators: exact control with constant (10%) display fault added, and constant (10%) control fault added with honest display. With the onset of a constant display fault (C_4 in Table 2.4), the level of human operator’s trust in the control remains virtually unchanged, while the level of trust in the display reduces dramatically (approximately 50%).

These results suggest that the human operators possibly were aware of the source of the faults introduced so that they could calibrate their trust levels easily. With the onset of a constant control fault (C_2 in Table 2.4) from the perfect condition, on the other hand, the level of human operator’s trust in the control reduces in half, from approximately 100 to 50, which shows operators’ good performance in calibrating their trust level. With the same condition (C_2), however, the level of human operator’s trust in the display also reduces from approximately 100 to 60. This situation implies either that the human operators were unable to detect the original source of the faults, which is the opposite of the analysis of the previous results on human operator’s level of trust behavior, or that the human operators just blame the display undeservedly. Thus, the human operators showed poor performance in calibrating their trust in a display when constant control faults are introduced.

Table 2.4 Selected Conditions Used in Muir and Moray (1996)’s Experiments

		Control Property	
		Exact	Constant Fault Added
		Perfect behavior	<i>Constant control fault added (10%) (C_2)</i>
Display Property	Honest		
	Constant Fault added	<i>Constant display fault added (10%) (C_4)</i>	Combination of two constant fault added

The results suggest that the human operators reduced their level of trust in a display even though the detected (or at least, presumed to be detected) faults came from an automation malfunction, not from a display malfunction. This characteristic of human behavior might be explained using the concept of “complacency” (Parasuraman et al., 1993). The concept of complacency asserts that the level of a human operator’s trust in automation, when viewed

objectively, is higher than it perhaps should be. Therefore, human operators do not always disengage and take over the automation when it goes wrong, but rather, still expect the automation to perform its function. This situation, however, is somewhat different from the conventional concept of complacency, in which the human operator's trust changes proportionally, if not appropriately. Not only did operators fail to calibrate their trust across two different subsystems, but also they were biased to lower their trust in only one system (displays), which could be characterized as the medium which presented the measured data to the human operators. In short, the human operators showed what could be called "hardware complacency" rather than "software complacency."

This result has two direct implications for IW studies. First, if the human operators regard the display unit as an independent agent performing as part of the whole process, this result might represent the human operator's ability to separate the data from the automation that produces the data. Hence, they can remain separate during further processing by human operators. Also, the result might imply that when the automation performs poorly, human operators tend to reduce their trust level, or get suspicious of the data communication system rather than the actual automatic machine performing the jobs. Since Muir and Moray (1996) did not analyze the situation, there is no statistical assurance for this assertion. However, based on the graphs reported in the paper, it is not difficult to see the difference.

Having completed our discussion of the characteristics of data communication between the automated system and human operators, we now turn our attention to a preventive method based on ecological interface design. According to Vicente and Rasmussen (1992), the degree of unfamiliarity or novelty can become a basis to classify events in complex human-automation interaction systems. They classified three anchor points along the familiarity continuum from the perspective of both human operators and designers.

1. Familiar events are routine in that human operators experience them frequently. As a result of a considerable amount of experience and training, human operators have acquired the skills required to deal with these events.
2. Unfamiliar but anticipated events occur infrequently, and thus, human operators will not have a great deal of experience on which to rely. However, the events have been anticipated by plant designers, who have built in means to deal with them (e.g., procedures, decision support systems, automatic controllers, etc.). These anticipated solutions provide human operators with the help they need to cope with this class of events.
3. Unfamiliar and unanticipated events are also unfamiliar to human operators because they rarely occur. Unlike the previous category, however, the event has not been anticipated by designers. Thus, human operators cannot rely on a built-in solution but must improvise one themselves.

Since it is important to detect abnormalities originated by the enemy force, designing an interface to overcome or minimize the effect of unfamiliar and unanticipated events becomes a critical issue. One way to accomplish this is using ecological interface design, which has been given much attention recently (Vicente, 1992a; 1992b; 1996). The concept of ecological

interface design has a common ground with other interface design philosophies. The fundamental difference from others is that it embraces decision making models which concentrate on the human behavioral and representational characteristics. The promise that these decision making models share is that humans tend to make decisions based on what they recognize from the environment (e.g., the RPD model), not based on what they think and deduce from the association between recognition and their mental models. Humans are apparently better at recognition-primed decision making than at deduction from mental models. Of course, this is not to say that humans depend solely on recognition for decision making. A few studies (e.g., Hammond, Hamm, Grassie & Pearson, 1987; Woods, 1988), however, have shown that human operators' performance was generally better when they relied on perceptual characteristics of the display than on functional properties of the process being controlled.

Another useful aspect of ecological interface design for the IW domain is that it can provide human operators with various viewpoints, in a hierarchical level, for the process or display being controlled. The number of viewpoints is totally domain-specific. Conceptually, complex systems are represented with a multiple level hierarchy, called an Abstraction Hierarchy (Rasmussen, 1985). Each level in the hierarchy consists of its own representation of the complex system. For example, Rasmussen (1985) found five levels of hierarchy to be useful for the description of the control systems being processed, as illustrated in Figure 2.2 and described below.

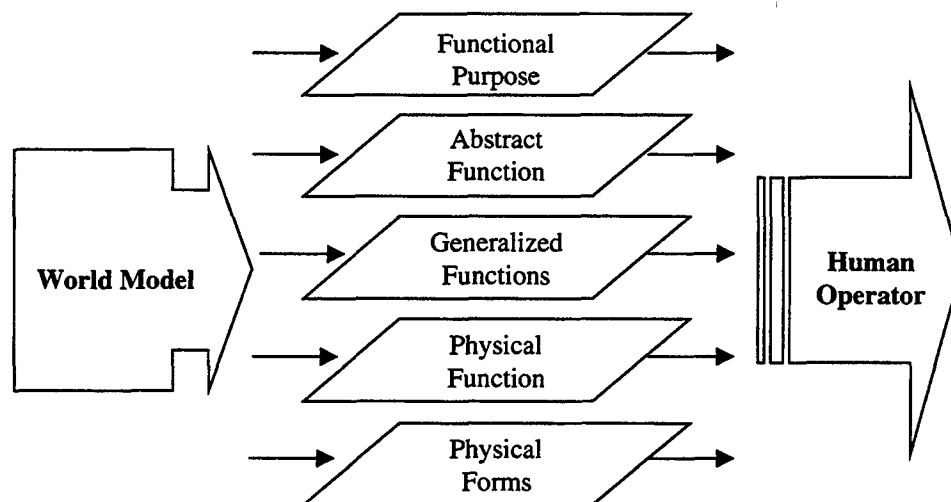


Figure 2.2 The system properties considered in human-machine interaction can be described at various levels of abstraction, representing the physical implementation and functional purpose in varying degrees.

1. *Functional Purpose*: Highest level of abstraction that translates the system's design purpose.
2. *Abstract Function*: Represents the intended causal structure of the process in terms of mass, energy, information, or value flows.
3. *Generalized Function*: Consists of the fundamental functions that the system is designed to achieve.

4. *Physical Function*: State of the system components that are designed to implement the generalized functions.
5. *Physical Form*: Physical instantiation of the system.

Each level represents its own understanding of the system being controlled. Having this abstraction hierarchy ready to assist them, human operators can more easily cope with unfamiliar and unanticipated events. In the previous section, we have discussed "Explication of Intention" as a characteristic of the "other actor." In short, the fact that humans can develop and calibrate their trust with ease when they communicate "well," also applies to the human-automation interaction environment. By this we mean that when we have various methods of communicating with automation, we should have better understanding, situation awareness, and manipulation of complex systems. Each level in the hierarchy, a way of representing the complex system, can match the human operators' level of intention, such as skill-, rule-, and knowledge-based behavior. Especially, the Abstraction Hierarchy is a way of supporting knowledge-based behavior. Generally, ecological interface design supports performance at all three levels of behavior—skill-, rule-, and knowledge-based—and use the Abstraction Hierarchy to support knowledge-based behavior in the face of unanticipated events (Vicente, 1990). Ashby's (1956) Law of Requisite Variety states that complex systems require at least equally complex controllers. Human operators cannot possibly control complex systems through a simplified interface design. In this perspective, the ecological interface design method will enable us to design complex systems for human operators to better understand and cope with the complex human-automation interaction environment.

2.3.3 Dynamics of Trust: Overtrust-Trust-Distrust-Mistrust

2.3.3.1 Calibration of Trust and Mistrust:

There are not equally competent machines; even different functions of a single machine are not equally competent, and certainly are not flawless. Thus, operators should learn how to adjust themselves within the given environment so that they neither distrust the good nor overtrust the poor quality of automation. Muir (1994) showed how an operator's trust and consequent selections of automatic or manual control mode interact with the quality of the automation to affect joint system performance.

Even though an operator's trust and the quality of the automated systems are described as each having two distinctive extreme states, these concepts provide us with a good understanding of the kinds of effects that occur when trust and the quality of automation are miscoupled. Interdependencies between informational quality and operator's views of trust and their allocations to or away from automation are shown in Table 2.5. Operators who calibrate their trust well (cells I, III), on the one hand, will know when to use and not to use automation; thus they will optimize joint system performance. Specifically, their appropriate trust, coupled with competent and reliable automation, will show them when and where to change their attention to compensate for the potential poor performance of less competent automation. Poorly calibrated operators (cells II, IV), on the other hand, tend to reject automation even when it shows good performance (cell IV), and also to accept poor automation (cell II). This situation may demand levels of operator

resources or abilities beyond those they are able to give, consequently preventing them from tackling required resource-intensive functions in the system and often causing “human errors.” Further, Muir (1989) suggested four possible ways to improve the calibration of trust toward decision support systems, in particular, see Muir (1989) for detailed discussion:

- Improve the user’s ability to perceive a decision aid’s trustworthiness
- Modify the user’s criterion of trustworthiness
- Enhance the user’s ability to allocate functions in a system
- Identify and selectively recalibrate the user on the dimension(s) of trust, which is (are) poorly calibrated.

Table 2.5 How the Operator’s Trust In and Use Of the Automation Interact with the Quality of the Automation to Influence System Performance. Adapted from Muir (1994).

Operator’s trust and allocation of function	Quality of the automation	
	‘Good’	‘Poor’
Trusts and uses the automation	(I) <i>Appropriate trust,</i> Optimize system performance	(II) <i>False trust,</i> Risk automated disaster
	(IV) <i>False distrust,</i> Lose benefits of automation, increase operator’s workload, risk human error	(III) <i>Appropriate distrust,</i> optimize system performance

Note: The cells illustrate appropriate trust, appropriate distrust, and the two errors of mistrust (false trust and false distrust).

Unfortunately, we seem far away from unblemished automation, which would have perfect reliability and performance, generating appropriate decision alternatives and no false alarms. False alarms have been especially acknowledged as a major contributor to the human operator’s miscalibration of trust. Current technology provides us with machines and computers which have relatively high degree of reliability. When false or inappropriate alarms occur, however, the human operators tend to become suspicious about the truthful states of the false alarms, which in consequence, might in the future be ignored and cancelled. This has been termed the “cry-wolf” phenomenon.

Sorkin and Woods (1985) provide a theoretical justification for the argument that high false alarm rates will have serious consequences on system performance and note that human operators of complex systems such as trains, aircraft (Bliss, 1997), and medical systems (Kerr, 1985) often turn off crucial alarm systems because of their tendency to activate without apparent reason. This phenomenon has a very important implication for the IW domain. In this domain, human operators must detect abnormalities or susceptibilities displayed on their screens. Having experienced false or inappropriate alarms associated with, for example, unfamiliar high-pitched sound warnings, they may regard similar future signals as repeated false alarms. In a sense, human operators would have calibrated their trust in automation (warning systems in this case), merely enough to acknowledge the presence of warning signals. When combined with warning systems, however, poorly calibrated human operator’s trust in complex systems is likely to create

complex human behavior. False alarms, generated by a defective warning system design representing the crucial states of complex systems make the human operators deactivate the warning system, consequently missing true signals.

2.3.3.2 Overtrust (Complacency):

On June 30, 1994, an Airbus A330 crashed during a test flight, killing seven on board. The purpose of the test flight was to evaluate the performance of the aircraft's autopilot system with an engine out, simulated hydraulic failures and unbalanced center of gravity just after takeoff. According to the investigating committee, the crew appeared overconfident and did not intervene in time to prevent the accident. The committee deduced that if the pilot had intervened and retaken manual control four seconds earlier than he actually did, the crash would have been avoided (Sparaco, 1994).

Overtrust in automation, sometimes referred to as "complacency" (Parasuraman, et al., 1993) has been scrutinized recently. Billings, Lauber, Funkhauser, Lyman and Huff (1976) defined complacency as "self-satisfaction which may result in non-vigilance behavior, based on an unjustified assumption of satisfactory system state." Like other causes of human errors, it reflects a mismatch between human capabilities and system/automation characteristics. In other words, complacency is a construct induced to elucidate operator behavior in interacting with a complex system and in failing to recognize that function(s) of the automation system have failed, or that it is in a different mode than the operator believes it to be.

A high level of trust in automation could lead operators to fail to vigilantly monitor their display and instruments, a state associated with reduced arousal. Analyses of aviation safety reporting systems have provided evidence of monitoring failures linked to excessive trust in, or over-reliance on, automated systems such as the autopilot and flight management system (Mosier, Skitka, & Korte, 1994; Singh, Molloy, & Parasuraman, 1992; Singh, Molloy, & Parasuraman, 1993a; Singh, Molloy, & Parasuraman, 1993b). Will (1991) also performed an experiment with reservoir petroleum engineers to determine the degree to which they relied on an expert system to perform well analysis. Using a modified expert information system, the engineers analyzed data on well pressure buildup problems and made decisions based on the expert system's recommendations. The system generated not only false recommendations, but also generated incorrect explanations to support them. The results showed that novices expressed higher confidence ratings in their decision making using the faulty expert system technology than those who drew wrong conclusions based upon the use of conventional hand-calculating methods. The phenomenon experienced by novices extended to experts. The results showed that experts were also deceived by the defective expert system, except in the case of one expert subject. While unaware that their own conclusions were wrong, the expert subjects said they thought they could have performed the jobs better without help from the expert system's decision aiding function. Verbal protocol analysis from the expert subjects (obtained through interviews after the experiment) revealed that they believed that they had sufficient knowledge about the process and the tasks, and therefore, refused to depend on the expert system. The experiment produced somewhat contradictory results, perhaps due to the small number of expert subjects recruited (five).

Then, what are the factors associated with this construct? Parasuraman and colleagues (Parasuraman, 1987, Parasuraman et al., 1993, 1996) have studied the factors influencing the monitoring automation, such as overall workload imposed on the operator. Under the single automation monitoring task, human operators performed very efficiently at detecting any automation failures, despite the fact that humans are very vulnerable to vigilance variation. Within a multiple task environment, however, performance is degraded (Parasuraman et al., 1993). Further, they examined the effect of consistency of automation reliability and concluded that human operators performed better at automation monitoring with inconsistent, variable reliability, because of the operator's lack of overreliance on automation due to the low level of trust. In addition to these factors, they addressed some others related to complacency:

1. **Information Overload:** Current technology tends to condense information presented to human operators, partly because of the economical reason, such as reduction of the number of human operators involved in a process. Often, human operators do not have time to monitor automation even if they overcome the limits of their monitoring capabilities.
2. **Information Underload:** Fused information reduces the amount of information human operators have to process. At the same time, however, the reduced performance pressure potentially can create a vigilance decrement associated with reduced arousal.
3. **Imperfect Mental Model:** Partly because of the fusion of information and the limited information it provides, human operators experience difficulty understanding the complex systems. With poor understanding of both the automation itself and the complex system it controls, human operators have no choice but to rely on automation, and hence are vulnerable when the automation fails. For instance, a typical complex system in current aircraft automation technology, the Flight Management System (FMS), often surprises pilots with unexpected mode transitions.
4. **Impaired Situation Awareness:** Having an imperfect mental model results in poor understanding of the current system state. Furthermore, it is probably impossible to estimate the future states of a complex system without knowing its current state. Thus, the ability to anticipate and predict future states of system, an aspect of situation awareness, will be impaired (Endsley, 1994).
5. **Diffusion of Responsibility:** Sheridan (1980) argued that "when authority and responsibility are shared, accountability becomes *diffuse*" (emphasis added). Consider the multiple quality control human inspectors on a single processing line. The second inspector probably will not inspect as vigorously as the first inspector, resulting in less visual inspection overall. Thus, the multiple agents involved in a system make each subsystem's roles and associated responsibilities ambiguous. Even when a small error develops into a disastrous one, it is difficult to find its fundamental source to blame.
6. **Cue utilization based on salience:** Weighting of information sources may reflect the ability to capture attention rather than the decision making value. Classical decision making theory based on utility emphasizes that cues have to be weighted based on their value toward total decision-making value to optimize a decision making, not based on salience. Unfortunately, people are often influenced by cue salience.

So far, we have been investigating how human operators develop their complacency on complex systems and the effects of such. It seems misleading that complacency is a psychological construct which human operators induce because of the limits of their capabilities on monitoring automation, information processing capacity and as such. However, system and automation characteristics also may result in a human operator's complacency. Lee, Parasuraman, and Bloomfield (1997) suggest:

1. **Observability:** By this, they mean when the system provides human operators with delayed or ambiguous feedback, the degree of human operators' observability decreases. In other words, the degree of transparency of system can contribute to the creation of the human operators' complacency.
2. **Time Urgency or Criticality:** Time available for human operators to make a decision and to take appropriate actions has a major impact on creating the human operator's complacency inducement. The less time the human operators have for a decision-making, the more opportunity it creates human operators' complacency.
3. **Automation Reliability:** Automation reliability determines the number of human operator interventions required to correct automation failures. Molloy and Parasuraman (1992) investigated human operator monitoring efficiency depending on automation reliability in detecting malfunctions and proved the commensurate relationship between the two measures. As automation reliability in detecting failures increases, the human operator's monitoring efficiency represented in the percentage time of manual performance decreases.
4. **Clumsy Automation:** Lee et al. (1997) argued that clumsy automation can induce complacency by the degree to which workload peaks are heightened and workload troughs are deepened.
5. **Complexity:** Because humans have both limited information processing and limited working memory capacity, complex work domains, in which the number of interacting elements exceeds the human operator's capacity, will induce operator complacency.
6. **Limited Functional Integration:** Humans must integrate the functions presented to them by automation through the automation interface. Thus various informational elements from and of automation require human integration.

Finally, we should point out an important aspect of overtrust in automation related to these factors. As mentioned above, the human operator's overtrust in an automated system can result in less monitoring activity, as well as more frequent engagement of the automatic rather than the manual control mode. The automatic control mode will keep human operators out of the automation loop which may reduce operator's understanding of the system (impaired mental model). Consequently, an extensive use of the automatic control mode can lead to degradation of human operator's skill (Drury, 1992). Therefore, without confidence in dealing with the automated systems, human operators are forced to rely more on the automation, which reinforces their overtrust on automation. Thus, it forms a vicious cycle, as shown in Figure 2.3. How then can we break this vicious cycle?

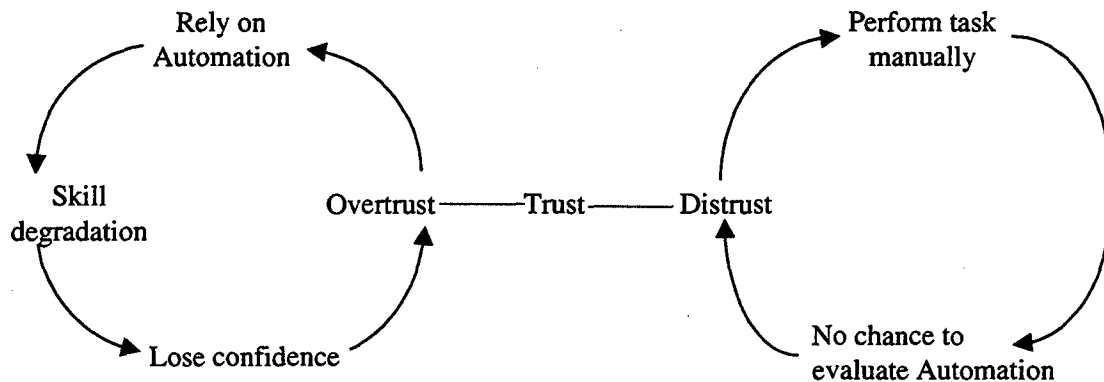


Figure 2.3 Two vicious cycles on the trust continuum

2.3.3.3 Distrust:

Having completed our discussion and critique of the characteristics of trust, calibration of trust and overtrust associated with various kinds of empirical data, we now turn our attention to the other end of trust continuum, distrust. If distrust merely is to define one extreme end of the trust continuum so as to be an opposite of trust (overtrust), it is not probably worth discussing. However, both ends could exist at the same time in human operators. A person can trust another partner on one characteristic while he/she does not trust the partner in other characteristics simultaneously. Distrust, therefore, is not a simple construct but might coexist with other trust elements. A common example for discussion of trust and distrust occurs in political matters, especially in presidential elections. We have witnessed many presidential elections and political disputes on the matter of a candidate's credibility or trustworthiness. For instance, Barber (1983) elaborated the criticality of trust's role in politics at length, by an example of the presidential election in '79-'80, and pointed out one of the main reasons why the Carter government lost the election was that it lost the public trust by blaming the public for its lack of confidence while it still managed to gain a high level of public's faith.

In human-computer interaction environments, where a human operator's main task is to monitor automation, novices tend to be biased toward distrust. A study by Riley (1996) illustrates this point. When two different subject groups, novices and experts, are given multiple tasks, experts tend to initiate automation more rapidly than novices when artificial failures are introduced into the tasks. Complete distrust of one component of automation may result in the human operator performing those tasks manually, a situation associated with a high level of workload and a decrease in overall performance (Moray & Lee, 1990; Lee & Moray, 1992). Also, disengaging automation and performing tasks manually leaves the human operators little opportunity to evaluate or reevaluate the automated system's trustworthiness, because the human operator necessarily reduces evidence of automation failure when he/she performs tasks manually. Consequently, the level of operator's distrust in automation remains intact. Again, it forms a second vicious cycle, in contrast to the cycle when the human operator overtrusts automation (see Figure 2.3).

2.4 Implications for IW

This section will review our knowledge of trust, make explicit the links to IW, and provide a model of trust to guide applications to IW.

2.4.1 What Do We Know About Trust?

First, we have established that the concept of trust is an internal state of the human operator, directly accessible only by the operator providing a scaling or judgment. Second, trust drives the operator's strategy in a task (e.g., the use of a manual vs. an automated system), which in turn drives operator and system performance. Thus, trust, although a subjective state, has important implications for performance in tasks where multiple strategies are possible (e.g., IW).

Third, the trust level can be changed by both system events and operator skills. In particular, trust can be reduced by error/failure/sabotage to the system. Thus trust, and hence system performance, are vulnerable to system malfunctions, caused by natural or by hostile-induced faults. Finally, the operator's level of trust does not always accord with objective truth. This can arise from known operator biases (e.g., poor sampling, complacency time lags during dynamic changes in trust, or poor estimation of own skills). Thus there are definite avenues of access to an operator's trust level, which could be exploited intentionally by an enemy in an IW context.

2.4.2 A Model of Trust for IW

Trust is an internal state of the human operator, but a state with proven links to antecedents within the operator and in the task, and proven effects on strategy, performance, and perhaps, operator well-being. Because of this structure, trust has same similarities to stress. Stress is an internal construct based upon the operators' perceptions of themselves and their tasks, both perceptions being influenced by, but not identical to, objective reality. Similarly, stress can drive the actions of the operator in terms of how to cope with the task. Finally, stress cannot be measured only with reference to externally verifiable phenomena; the operator's perception is a required ingredient.

We have seen that the level of trust may not reflect situational realities, and we can model this aspect rather easily. As Luhman (1980) indicates, trust is a basic fact of social life and is the very necessary element to "reduce complexity" in social life. He further explains that "the world presents itself as unmanageable complexity, and it is this [trust] which constitutes the problem for systems which seek to maintain themselves in the world." Thus, trust can be seen as that which reduces apparent complexity; mismatches between the level of trust operators should exhibit and the level they do exhibit come from two sources:

- Mistrust (i.e., the operators do not trust as much as they should)
- Overtrust (i.e., the operators trust more than they should)

A convenient analogy, linking mistrust and overtrust to the flow of trust from a situation to the operator, is the flow of information along a transmission channel (see Fig. 2.4).

Here, T_{in} is the true amount of trust which an objective and all seeing observer would find in the situation. Some of this, T_{lost} , does not reach the operator and is lost. Examples are monitoring failure, inadequate sampling, or mistrust due to disguised enemy action. Additionally, some overtrust (T_{add}) may be added by the operator, although it has no basis in reality. Sampling biases, overconfidence in the integrity of an algorithm, or overtrust due to enemy interference would all be examples of falsely added trust.

The outcome is the trust perceived by the operator, T_{out} . This may be larger or smaller than T_{in} , depending upon the relative magnitude of T_{lost} and T_{add} . But only a fraction of the operator's trust, T_{out} , is in fact justified by the situation, T_{in} . This amount of shared trust, here called transmitted trust, or T_{trans} , is the amount common to input and output, analogous to the information transmitted correctly from source to destination.

Note that

$$T_{in} = T_{trans} + T_{lost}$$

$$T_{out} = T_{trans} + T_{add}$$

Thus,

$$T_{trans} = (T_{in} + T_{out} - T_{lost} - T_{add}) / 2$$

From this model we can begin to locate and categorize the sources of trust, mistrust and distrust for a given system and operator. In the end, we would hope to be able to relate these sources to some of the dimensions of trust (e.g., Rempel et al.'s [1985] list of predictability, dependability and faith). In an IW context we should be able to explore systematically the effects of system changes on the levels of trust measured. It is our belief that an adequate model of the trust concept is necessary if we are ever to predict operator's actions, strategy, and performance from the antecedent conditions known to affect them.

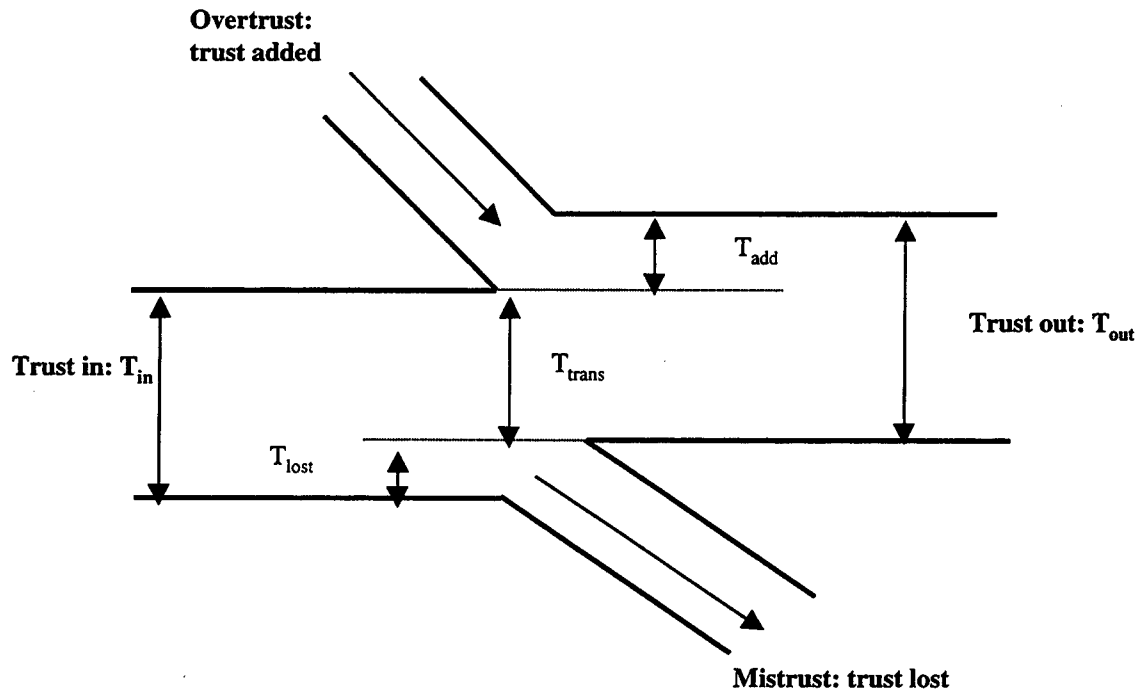


Figure 2.4 Trust transmission model

However, the model in Figure 2.4 is at too high a level of abstraction to guide the search for the structural element of the trust concept, and hence, the ways in which trust can be defended or attacked in an IW context. For this, we need to explore the way in which the operator estimates trust (i.e., how does the current value of T_{out} arise) and how well it reflects T_{in} . An obvious model here is one which examines the cues used by the operator in forming trust and how well these cues reflect the true situation. This is the Lens model (Brunswik, 1952), shown in Figure 2.5 as modified later by Cooksey (1996). Brunswik's Lens model is a symmetrical framework which describes how both the *environmental structure* and patterns of *cue utilization* collectively contribute to judgment performance. In this model, the judge combines cue information (X_i) about the environment to make a judgment (Y_s). The model represents the classical notion of information transformation from stimulus (information presentation) to response (judgment) in which humans process information internally to yield some functional response based on the cues observed, which in turn, are representations of the environmental state. Thus, the model includes not only a classical decision concept (i.e., how humans sample and combine the cues presented to them) but also the relationship between available cues and the true state of the environment.

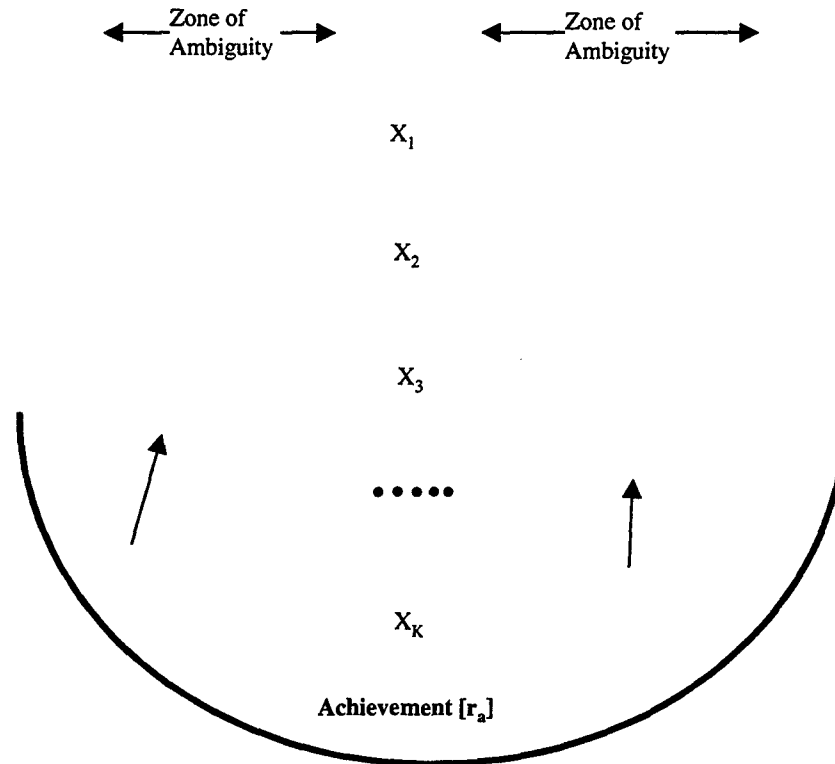


Figure 2.5 Brunswik's Lens Model adapted from Cooksey (1996).

By analyzing a judge's cue utilization policy, therefore, we may be able to understand how that judge has adapted to the structure of the environment. The predictability of the environment, given a set of cues (the ecological validity of the cues) can also be assessed. Therefore, this model allows us to assess and evaluate how well the true environment structure is represented via a set of cues. Additionally, achievement, denoted as r_a , represents how well human judgments correspond to the actual values of the environmental criterion to be judged. Achievement is shown in Figure 2.5 as a line connecting judgments to criterion values. Because the Lens Model provides the means for considering the judge's adaptation to the environment, and the degree of achievement, both of which relate to the calibration of human trust, it seems that the use of the Lens Model approach to model human trust in automated systems is reasonable.

2.4.3 Applying the Lens Model to Human Trust in Complex, Automated Systems

Conceptually, modeling human trust in automated systems using the Lens Model, shown in Figure 2.6, is relatively straightforward (Seong & Bisantz, in press). The judgment modeled in this case is the operator's judgment of the trustworthiness of some system component or output. That is, the operator decides whether or not a system component is to be trusted. In Lens Model terms, then, the environmental criterion is the actual trustworthiness of the component. The judgment is the operator's assessment of that trustworthiness. To make this judgment, the operator must rely on a set of observable cues which have some relationship to the components' trustworthiness. In this paradigm, the concept of calibration is explicitly measured by

achievement (r_a)—the extent to which the operator's assessment of trustworthiness matches the true state of the environment. One can also consider calibration to include the operator's adaptation to the structure of the environment, in terms of the relationship between the cues and actual integrity of information.

Further specification and experimental verification of this model of trust in automation beyond the general level noted above presents certain difficulties, however. First, there is no clear, objective measurement of the true state of environment in terms of its trustworthiness. Generally, trust as a state in itself has been measured only subjectively. This is problematic in terms of the Lens Model formulation, since application of the Lens Model and evaluation of the model parameters requires knowledge of the *true* environmental state. To circumvent this difficulty, we propose transforming the judgment from one of an assessment of trustworthiness to one that is more performance-oriented. From an engineering standpoint, we are interested in human trust in a system to the extent to which that trust affects system performance. For instance, we are interested in whether or not operators utilize an automated controller or obtain certain data, given their trust in that controller or information source. The true state of the environment, in terms of the adequacy of the controller or the integrity of the data source, can be objectively determined. For these examples, the operator's judgment would be whether to use the controller or the data. More generally, the operator's judgment is one of component utilization, and the true state of the environmental criterion is whether or not the component should have been used. In terms of trust, this assumes that an operator's behavior in utilizing a system component reflects their trust in that component.

Second, to implement a Lens Model description of human trust in automation, it is necessary to specify what cues might be available for an operator to make a judgment about whether to use a system component. Candidate cues include the components of trust identified by previous studies of trust (e.g., Barber, 1983; Rempel, et al., 1985; Zuboff, 1988). For instance, cues could include such factors as predictability, dependability, faith, reliability, competence or robustness. To be included in a quantitative Lens Model, these cues would be both measurable and available to the operator. The availability of these candidate cues to the operator depends to some extent on how information is displayed to operators. However, the consideration of how to measure these cues must be addressed. For example, consider predictability. If we define the environment to be judged in terms of a subsystem or set of systems, we can represent predictability in terms of the degrees of freedom in performance that were designed into the system. That is, predictability could be measured in terms of allowed error or performance variance. The smaller the degree of freedom, or allowable error, the more predictable the system is. If predictability is one component of trust, as Barber claimed, then trust will be negatively impacted by a large degree of performance variability. Additionally, the reliability of a system or component could be measured in terms of past performance (e.g., breakdowns, errors, etc.).

2.4.4 Instantiating the Model

To evaluate the model, an experimental framework has been established in an IW domain (Seong, Llinas, Drury, & Bisantz, in press) in which one can consider trust in the context of aided

adversarial decision making, where military officers must assess the integrity of information which may be intentionally altered or degraded by an enemy. In this domain, the points of attack by an enemy can be the real battle situation, data gathering or fusion algorithms, or a data transfer network. By changing the points of simulated attack, we may be able to observe how operators successfully calibrate their trust in terms of accurately pinpointing the point of attack and changing the level of trust. In the IW domain, studying human trust is important for several reasons. For example, forces might be vulnerable to information attacks which diminish their trust in DF or other decision aids, rendering these assets less useful, or to deceptive attacks, in which an inappropriately high level of trust in the aid is maintained. In terms of the Lens Model approach, data (fusion algorithm outputs) would be judged as usable or not (e.g., trustworthy or not), based on operators' understanding of the predictability, reliability, etc., of the information displayed to them.

A Lens Model approach for modeling human trust in automated systems has been proposed. Because the Lens Model provides the means for modeling both human judgment policy and the actual structure of the environment, it allows operator calibration to the actual trustworthiness of a system to be explicitly considered. Conceptual solutions for addressing certain difficulties with this approach, such as the objective determination of the true state of system trustworthiness and the identification and measure of cues which reflect system trustworthiness, were discussed. Finally, an experimental framework in the domain of IW was described which may provide the means for further instantiating and evaluating the effectiveness of this model of human trust in automation.

We have now augmented the trust transmission model by postulating cue use as an intervening variable. Thus each observable characteristic can contribute to transmitted trust, lost trust and added trust. The overall performance level $[r_a]$ corresponds to the transmitted trust, but the Lens model postulates mechanisms by which it is developed. Note that the trust transmission model is value-free, in the sense that we only consider the correspondence between input and output trust. In contrast, the Lens model makes values explicit (i.e., the values to the operator of the various cues).

The nature of the cues themselves can be derived from our knowledge of trust measures (Section 3) and factors affecting trust. For example, the dictionary definitions, the various trust frameworks (Section 2.2), and Sheridan's (1988) list of trust characteristics can all contribute potential cues. We could eventually see a mixed list, which includes reliability, robustness, and familiarity from Sheridan's work on trust and observability and complexity from the complacency list of Lee, Parasuraman, and Bloomfield (1997). Other cues could well be added as our knowledge of the measurement of trust develops. Similarly, our knowledge of human error causes (e.g., biased sampling, confirmation bias, and complacency) can lead to insights into cue utilization by operators.

In summary, our review of the twin literatures on human-human trust and human-automation trust has led us to models of trust transmission and trust estimation, both of which we believe are highly applicable in an IW environment.

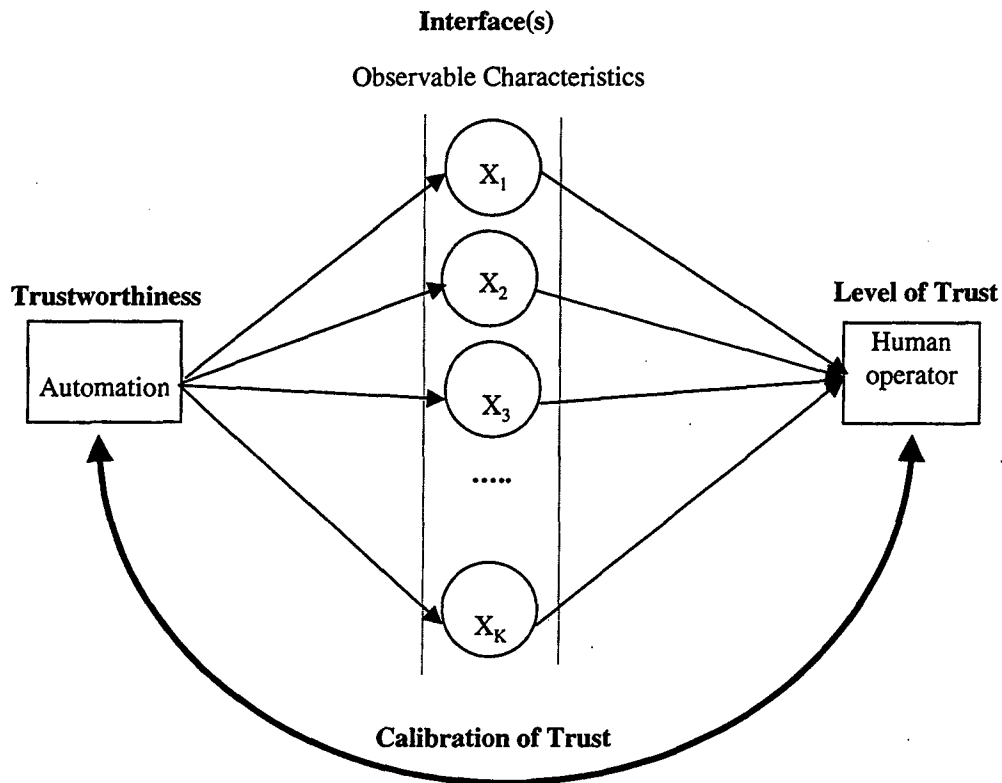


Figure 2.6 Model of human trust in automation using the Lens model.

2.5 References

- Ashby, W. R. (1956). *An introduction to cybernetics*. London: Methuen.
- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19, 775-779.
- Barber, B. (1983). *The logic and limits of trust*, New Brunswick, NJ: Rutgers University Press.
- Beth, T., Borcharding, M., & Klein, B. (1994). Valuation of trust in open networks. In D. Gollman (Ed.), *Computer Security—ESORICS 94, Third European Symposium on research in Computer Security*, (pp. 3-18). Brighton, UK: Springer-Verlag.
- Billings, C. E., Lauber, J. K., Funkhouser, H., Lyman, E. G. & Huff, E. M. (1976). *NASA aviation safety reporting system quarterly report* (Report No. 76-1). Moffett Field, CA: Ames Research Center.
- Bliss, J. P. (1997). Alarm reaction patterns by pilots as a function of reaction modality. *The International Journal of Aviation Psychology*, 7(1), 1-14.

- Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago, IL: Univ. Chicago Press.
- Christianson, B. & Harbison, W. S. (1997). Why isn't trust transitive? In M. Lomas (Ed.), *Security Protocols in Proceedings of the International Workshop* (pp. 171-176). Cambridge, UK: Springer.
- Cooksey, R. W. (1996). *Judgment analysis: theory, methods and applications*. New York: Academic Press.
- Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution*, 2(4), 265-279.
- Deutsch, M. (1960). Trust, trustworthiness, and the F scale. *Journal of Abnormal and Social Psychology*, 61(1), 138-140.
- Drury, C. G. (1992). Automation. In Federal Aviation Agency's *Human factors guide for aviation maintenance*. Washington, DC: U.S. Department of Transportation, Federal Aviation Agency.
- Endsley, M. (1994). Situation awareness in dynamic human decision making: Theory. In R. D. Gilson, D. J. Garland, and J. M. Koonce (Eds.), *Situation awareness in complex systems. Proceedings of a CAHFA [Center for Applied Human Factors in Aviation] Conference* (pp. 27-58). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.
- Hammond, K. R., Hamm, R. M., Grassia, J. & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Mans, and Cybernetics*, SMC-17, 5, 753-770.
- Holmes, J. G. (1991). Trust and the appraisal process in close relationships. *Advances in Personal Relationships*, Vol. 2, (pp. 57-104). London: Jessica Kingsley.
- Jones, S. & Marsh, S. (1997). Human-computer-human interaction: Trust in CSCW. *SIGCHI Bulletin*, 29(3), 36-40.
- Kantowitz, B. H. & Campbell, J. L. (1996). Pilot workload and flightdeck automation. In R. Parasuraman and M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 117-136). Mahwah, NJ: Lawrence Erlbaum.
- Kerr, J. H. (1985). Auditory warnings in intensive care units and operating theatres. *Ergonomic International*, 85, 172-174.

- Klein, G. (1997). The Recognition-Primed Decision (RPD) model: Looking back, looking forward. In C. E. Zsombok and G. Klein (Eds.), *Naturalistic decision making* (pp. 285-292). Mahwah, NJ: Lawrence Erlbaum.
- Larzelere, R. E. & Huston, T. L. (1980, August). The dyadic trust scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and the Family*, 42(3), 595-604.
- Lee, J. D. (1992). *Trust, self confidence, and operator's adaptation to automation*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.
- Lee, J. D. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243-1270.
- Lee, J. D. & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153-184.
- Lee, J. D., Parasuraman, R. & Bloomfield, J. R. (1997). A Multimodal Perspective on "Automation-Induced Complacency." A workshop presented at the Human Factors and Ergonomics Society 41st Annual Meeting, Albuquerque, NM.
- Lerch, F. J. & Prietula, M. J. (1989). How do we trust machine advice? In G. Salvendy and M. J. Smith (Eds.), *Designing and using human-computer interfaces and knowledge based systems* (pp. 410-419). Amsterdam: Elsevier.
- Levis, A. H., Moray, N. & Hu, B. (1994). Task decomposition and allocation problems and discrete event systems. *Automatica*, 30(2), 203-216.
- Luhman, N. (1980). *Trust and power*. New York: John Wiley.
- McClellan, J. M. (1994, June). Can you trust autopilot? *Flying*, 76-83.
- McMaster, R. C., McIntire, P. & Mester, M. L. (1986). *Nondestructive Testing Handbook, Vol. 4, Electromagnetic testing: Eddy current, Flux leakage and Microwave Nondestructive testing*, (2nd ed.). Columbus, OH: American Society for Nondestructive Testing.
- Moffa, A. J. & Stokes, A. F. (1996). Trust in a medical system: Can we generalize between domains? In M. Mouloua and J. M. Koonce (Eds.), *Human-automation interaction: Research and practice* (pp. 218-224). Mahwah, NJ: Lawrence Erlbaum.

- Molloy, R. & Parasuraman, R. (1992). Monitoring automation failures: Effects of automation reliability and task complexity. *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting*, 1518-1521.
- Moray, N. & Lee, J. D. (1990). Trust, automation and performance in human-machine systems. In W. Karwowski and M. Rahimi (Eds.), *Ergonomics of hybrid automated systems II*, (pp. 669-673). Amsterdam: Elsevier.
- Mosier, K. L., Skitka, L. J. & Korte, K. J. (1994). Cognitive and social psychological issues in flight crew/automation interaction. In M. Mouloua and R. Parasuraman (Eds.), *Human performance in automated systems: Current research and trends*. Hillsdale, NJ: Lawrence Erlbaum.
- Muir, B. M. (1989). Operator's trust in and use of automatic controllers in a supervisory process control task. Unpublished doctoral dissertation, University of Toronto, Canada.
- Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11), 1905-1922.
- Muir, B. M. & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429-460.
- Parasuraman, R. (1987). Human-computer monitoring. *Human Factors*, 29(6), 695-706.
- Parasuraman, R., Molloy, R. & Singh, I. L. (1993). Performance consequences of automation-induced "complacency." *The International Journal of Aviation Psychology*, 3(1), 1-23.
- Parasuraman, R., Molloy, R., Mouloua, M. & Hilburn, B. (1996). Monitoring of automated systems. In R. Parasuraman and M. Mouloua (Eds.), *Automation and human performance: Theory and applications*, (pp. 91-116). Mahwah, NJ: Lawrence Erlbaum.
- Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2), 230-253.
- Pulford, B. D. & Colman, A. M. (1996). Overconfidence, base rates and outcome positivity/negativity of predicted events. *British Journal of Psychology*, 87(3), 431-445.
- Rasmussen, J. (1983). Skills, rules, knowledge: Signals, signs, and symbols and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13, 257-267.

- Rasmussen, J. (1985). The role of hierarchical knowledge representation in decision-making and system management. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-15* (2), 234-243.
- Rempel, J. K. & Holmes, J. G. (1986, February). How do I trust thee? *Psychology Today*, 28-34.
- Rempel, J. K., Holmes, J. G. & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, 49(1), 95-112.
- Riley, V. (1996). Operator reliance on automation: Theory and data. In R. Parasuraman and M. Mouloua (Eds.), *Automation and human performance: Theory and applications*, (pp. 19-35). Mahwah, NJ: Lawrence Erlbaum.
- Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35, 651-665.
- Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist*, 26, 443-452.
- Scanzoni, J. (1979). Social exchange and behavioral interdependence. In R. L. Burgess, and T. L. Huston (Eds.), *Social exchange in developing relationships* (pp.61-98), New York: Academic Press.
- Seong, Y., Bisantz, A. M. (in press). Modeling human trust in complex, automated systems using a Lens model approach. K. Krahel & M. W. Scerbo (Eds.), *Automation technology and human performance: Current research and trends*. Mahwah, NJ: Lawrence Erlbaum.
- Seong, Y., Llinas, J., Drury, C. G., & Bisantz, A. M. (in press). Human trust in aided adversarial decision-making systems. K. Krahel & M. W. Scerbo (Eds.) *Automation technology and human performance: Current research and trends*. Mahwah, NJ: Lawrence Erlbaum.
- Sheridan, T. B. (1980, October). Computer control and human alienation. *Technology Review*, 61-73.
- Sheridan, T. B. (1988). Trustworthiness of command and control systems. *IFAC Man-Machine Systems*, 427-431.
- Sheridan, T. B. (1992). *Telerobotics, automation and human supervisory control*. Cambridge, MA: MIT Press.

- Sheridan, T. B. (1997). Supervisory control. In G. Salvendy (Ed.), *Human factors and ergonomics handbook* (pp. 1295-1327). New York; Wiley.
- Sheridan, T. B. & Johanssen, G. (1976). *Monitoring behavior and supervisory control*. New York: Plenum.
- Sheridan, T. B., Vamos, T. & Aida, S. (1983). Adapting automation to man, culture and society. *Automatica*, 19(6), 605-612.
- Singh, I. L., Molloy, R. & Parasuraman, R. (1992). Development and validation of a scale of automation-induced "complacency." *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting*, 22-25.
- Singh, I. L., Molloy, R., & Parasuraman, R. (1993a). Individual differences in monitoring failures of automation. *The Journal of General Psychology*, 120(3), 357-373.
- Singh, I. L., Molloy, R., & Parasuraman, R. (1993b). Automation-induced "complacency." Development of the complacency-potential rating scale. *The International Journal of Aviation Psychology*, 3(2), 111-122.
- Sorkin, R. D. & Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human Computer Interaction*, 1, 49-75.
- Sparaco, P. (1994). A-330 crash to spur changes at Airbus. *Aviation Week and Space Technology*, 141(6), 20-22.
- Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Vicente, K. J. (1990). A few implications of an ecological approach to human factors. *Human Factors Bulletin*, 33(11), 1-4.
- Vicente, K. J. (1992a). Multilevel interfaces for power plant control rooms I: An integrative review. *Nuclear Safety*, 33(3), 381-397.
- Vicente, K. J. (1992b). Multilevel interfaces for power plant control rooms II: A preliminary design space. *Nuclear Safety*, 33(4), 543-548.
- Vicente, K. J. (1996). Improving dynamic decision making in complex systems through ecological interface design: A research overview. *System Dynamics Review*, 12(4), 251-279.

Vicente, K. J. & Rasmussen, J. (1992). Ecological Interface Design: Theoretical Foundations. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-22 (4), 589-606.

Webster's third new international dictionary of the English language, unabridged. (1993). Springfield, MA: Merriam-Webster.

Will, R. P. (1991). True and false dependence on technology: Evaluation with an expert system. *Computers in Human Behavior*, 7, 171-183.

Woods, D. D. (1988). Coping with complexity: The psychology of human behavior in complex systems. In L. P. Goodstein, H. B. Anderson, and S. E. Olsen (Eds.), *Tasks, errors and mental models* (pp. 128-148). New York: Taylor & Francis.

Zuboff, S. (1988). *In the age of the smart machine: The Future of work and power.* New York: Basic Books.

3.0 MEASURES OF TRUST AND RELATED NOTIONS

3.1 Overview

Previous research in the area of trust has used two basic types of measures for assessing trust in different situations: personal rating measures and process and performance measures. Rating techniques, such as questionnaires, may seem to be the most direct way to measure a subjective feeling or state of mind (such as the amount of trust one has in another person) or in the information provided in a complex system. However, more objective measures can also provide information relevant to the assessment of trust. For instance, people's decision making or control strategies may change based on the trust they have in the available information, or the quality of various systems components. The following sections summarize rating measures that have been used in previous studies. Section 3.2 describes rating measures of trust that have been used in previous research. Section 3.3 outlines an approach for systematically developing a scale for measuring feelings of trust between humans and systems and presents preliminary results from a study based on this approach. Finally, Section 3.5 describes human-system process and performance measures of trust that have been used in previous systems and how these can be applied to the measure of trust in an IW environment.

3.2 Rating Measures of Trust

One method by which trust has been assessed is through the use of questionnaires, or rating scales. These subjective trust scales include the *Trust Scale* (Rempel, Holmes & Zanna, 1985, Rempel & Holmes, 1986), the *Interpersonal Relationship Scale* (Rempel et al, 1985; Rotter, 1967, 1980), and the *Dyadic Trust Scale* (Larzelere & Huston, 1980). In general, these questionnaires were developed from a social science perspective, and focused on investigating different aspects of trust between humans in particular. No consistent set of characteristics were studied across the questionnaires. Subjective measures of trust have also been used in conjunction with experimental studies in which participants are asked to rate their trust in aspects of a *system* they were controlling (e.g., Lee & Moray, 1994; Muir & Moray, 1996). Additionally, an initial study of the dimensions or features of automation-related complacency (signifying overtrust) was conducted by Singh, Molloy and Parasuraman (1993a). The following sections summarize the content and findings of some previous studies which used subjective measures of trust.

Trust Scale. Rempel et al. (1985) developed a trust scale to measure what they hypothesized to be three independent components of trust: predictability, dependability, and faith. The concept of trust can be considered a construct with a number of different elements, each contributing to the overall feeling of trust. Their study was based on the notion that people attempt to understand their partners in terms of acts, dispositions, and motives that would predict positive responses (Rempel et al., 1985). The authors related the recommended constructs of trust both to each other and to feelings of love and satisfaction within a close relationship. Twenty-six topical questions were developed and refined in this scale, derived from earlier scale-related studies such as Rubin's Loving and Liking Scale (Rempel et al., 1985), and were designed to measure levels of trust within close interpersonal relationships (see Table 3.1 for

selected example questions). In a later study, Rempel and Holmes (1986) constructed a similar, but more condensed trust scale. In general, Rempel and his colleagues found that trust is related in important ways to the success of a close relationship. Questions on the Trust Scale addressed multiple hypothesized dimensions of trust, such as faith, dependability, and predictability.

Table 3.1 Selected Questions from Rempel et al.'s (1985) Trust Scale. Respondents Were Asked to Rate Their Agreement with the Statements, Using a Seven-Point Scale.

	Question	Designated Trust Category
1	When we encounter difficult and unfamiliar new circumstances I would not feel worried or threatened by letting my partner do what he/she wanted.	F
2	I can count on my partner to be concerned about my welfare.	D
3	In general, my partner does things in a variety of different ways. He/she almost never sticks to one way of doing things.	P

* F=faith; D=dependability; P=predictability

Interpersonal Relationship Scale. Rotter (1967) constructed a different measurement scale for trust, called the Interpersonal Trust Scale. For this study, trust was defined as an expectancy, held by an individual or a group, that *other people can be believed*. Example questions used as a motivating framework for rating trust, using Rotter's Interpersonal Relationship Scale, are shown in Table 3.2. In contrast to Rempel et al. (1985), the scale was intended to measure *general* interpersonal trust rather than trust in a *specific* relationship.

Table 3.2 Example Questions from the Interpersonal Relationship Scale (Rotter, 1967). Participants Were Asked to Rate Their Agreement with the Statements on a Five Point Scale from Strongly Agree to Strongly Disagree.

Parents usually can be relied upon to keep their promises.
Hypocrisy is on the increase in our society
In dealing with strangers, one is better off to be cautious until they have provided evidence that they are trustworthy.

Dyadic Trust Scale. Larzelere and Huston (1980) developed the Dyadic Trust Scale, which was intended to measure interpersonal trust between romantically linked partners, or *dyads*. The authors defined trust as the extent to which a person believes another person (or persons) to be *benevolent and honest*. Larzelere and Huston (1980) used factor analysis to identify eight independent components of trust, based on an initial set of 57 items, which were adapted from previous scales (e.g., Rotter, 1971, Schlenker, Helm, & Tedeschi, 1973). Example questionnaire items are shown in Table 3.3. Responses from this scale correlated highly with scales of love but did not correlate significantly with measures of generalized trust, indicating that the scale measured interpersonal rather than general feelings of trust.

Table 3.3 Example Statements from the Dyadic Trust Scale Study (Larzelere & Huston, 1980). Participants Were Asked to Rate Their Agreement with the Statements on a Seven-Point Scale from Strongly Agree to Strongly Disagree.

My partner is primarily interested in his/her own welfare.
My partner is perfectly honest and truthful with me.
My partner treats me fairly and justly.

Complacency Potential Rating Scale. Feelings of complacency are related to feelings of trust and possibly overtrust. For instance, people are more likely to behave complacently, feeling secure in their awareness of a situation, if they trust the information they are seeing, or the ability of a system to perform according to expectations. However, a false sense of complacency may result in an undesirable outcome. For example, complacency is one of the behavioral coding categories used to classify aircraft flying incidents in the Aviation Safety Reporting System (ASRS) (Singh et al., 1993a). A feeling of complacency may result in non-vigilance based on an unjustified assumption of satisfactory system state. To study this issue, Singh et al. developed and evaluated a rating scale to measure someone's *potential for becoming complacent with automated technology*. The study focused on investigating attitudes towards everyday automated devices such as automated teller machines. They claim that people who show more trust in and reliance on automation will have a higher potential for complacency. Example motivational statements from this rating scale, which contained 20 items, can be found in Table 3.4. Factor analysis was used to identify factors contributing to the overall complacency potential score, including general attitudes toward automation, and confidence-related, reliance-related, trust-related, and safety-related attitudes toward automation.

Table 3.4 Example Motivational Statements from the Complacency Potential Rating Scale Study (Singh et al., 1993). Participants Were Asked to Rate Their Agreement with the Items, Using a Five-Point Scale Ranging from Strongly Agree to Strongly Disagree.

Manually sorting through card catalogues is more reliable than computer-aided searches for finding items in a library.
If I need to have a tumor in my body removed, I would choose to undergo computer-aided surgery using laser technology because computerized surgery is more reliable and safer than manual surgery.
People save time by using automatic teller machines (ATMs) rather than a bank teller for banking transactions.

Trust in Automated Systems. Lee and Moray (1994) and Muir and Moray (1996) performed experiments using process control simulations, in which they manipulated the quality or various system components. In addition to several performance and process measures, they had participants rate their trust in different system aspects using a subjective scale, shown in Table 3.5. (See also the discussion on these studies in Section 2.)

Table 3.5 Motivational Statements from the Subjective Rating Scale Study of Lee and Moray (1994).

Trust in the local bus service to get you to the store on time./Self-confidence in your ability to get to the store in time.
Trust in your calculator or computer to produce the right answer./Your self-confidence in your ability to arrive at the correct answer doing the calculations manually.
Trust in the heating system where you live to keep you comfortable./Your self-confidence in your ability to turn the heater on and off manually to keep you comfortable.
Trust in your watch to tell the correct time./Your self-confidence in your ability to estimate the correct time.

Trust in Advice Giving Systems. Lerch and Prietula (1989) investigated the effects of the source of financial management problem solving advice on self-reported measures of agreement with the advice and confidence, or trust, in the source of the advice. Advice provided to participants was attributed to either expert systems (i.e., to automated decision aids) or humans with different qualifications (expert vs. novice). Participants were asked to rate their agreement with the advice and their confidence in the source of the advice. Confidence was assessed at the outset of the experiment, reflecting any differences in participants' a priori attributions of dependability of the advice source. Subsequently, confidence was assessed after participants received advice, solved problems, and received feedback about the appropriateness of the advice. Lerch and Prietula hypothesized that these subsequent measures of trust would reflect how trust changes based on the quality of the advice. The authors found that, overall, participants were less confident in expert systems and human novices than human experts. Additionally, trust, as measured by a subjective rating of confidence, in all three sources changed as the quality of the advice, and thus participants' agreement with the advice, changed. Trust increased after participants agreed with advice on some problems (which was typically good advice) and declined after participants showed less agreement with advice on other problems (which was typically poor advice), indicating that participants were updating their initial ratings of confidence by considering their experience with the advice.

3.3 Developing an Empirically Based Scale to Measure Trust

People's feelings of trust have been measured directly using many different types of questionnaires, in both social psychology and engineering, as described above. However, the questionnaires used to measure trust in both social psychology and human-machine systems research have not been based on an empirical analysis and determination of what kinds of factors or items should be included in the questionnaire. Instead, each questionnaire was designed based on the researcher's theories about the meaning of trust, and the multiple components of trust within these theories. Additionally, the previous studies have not explicitly evaluated how trust between human and automated systems differs from trust between humans, or for that matter from trust in general. This is important, because any trust measurement scale used to evaluate trust in automated systems such as those used in an AADM environment should be based on those factors important to trust between humans and systems, rather than factors of trust between people, or trust in general.

Since the concepts investigated in different questionnaires for different research purposes may or may not be independent of each other, simply creating a comprehensive questionnaire by combining items from previous questionnaires would not be appropriate. To develop a descriptive model of human trust in any situation, we need a questionnaire that is comprehensive and measures a set of factors which contribute independently to overall trust.

To identify potential similarities and differences among concepts of general trust, trust between people, and trust between humans and systems, and between trust and distrust, we are currently performing a three-phased, empirical study, modeled after that used by Zhang, Helander, and Drury (1996). In the first phase, we collected various words related to concepts of trust and distrust. In the second phase, we investigated how closely each of these words was related to trust or distrust in order to evaluate whether or not trust and distrust were opposites or represented completely different concepts, and whether or not concepts of trust and distrust were similar for general trust, trust between people, and trust between humans and systems. In the third phase, we are conducting a paired comparison study to identify multiple, independent factors of trust and distrust.

Phase 1

Objective. The objective of this phase was to collect a large set of words related to trust and distrust.

Participants. Seven students majoring in Linguistics or English were recruited, because of their presumed knowledge of word meanings. Our goal was to identify as many words as possible which were related to trust and distrust (for all three categories of general, human-automation, or human-human trust), for use in subsequent experiments. All participants were native English speakers.

Procedure. There were three conditions in this experimental phase. Participants were asked to provide written descriptions of their understanding of both trust and distrust with respect to either trust between people, trust in automation, or trust with no further qualification. Additionally, an initial set of 138 words was collected by analyzing questionnaires used in previous studies, and from dictionary definitions and thesauri. Participants were asked to rate whether those words were related to trust using a nominal scale, with "positively related to trust," "not related to trust," "negatively related to trust," and "don't know." As with the written descriptions, these ratings were done with respect to the three conditions of trust between people, trust in automation, and general trust.

Results. We obtained 36 new words from the written descriptions of trust provided by participants. Additionally, we eliminated words from the initial set of 138 words based on participants' ratings of the words. Words which were rated "not related to trust" by four or more participants were eliminated. We also eliminated words that were ambiguous: that is, words which some participants rated as "positively related to trust" while other participants rated as "negatively related to trust." The final set of words, which we will refer to as Set-1, contained 112 trust-related words.

Phase 2

Objective. The objective of this phase was first, to determine whether the concepts of trust and distrust are inversely related; and second, to determine whether concepts of trust and distrust are similar across general, human-human, and human-machine trust.

Procedure. In this experiment, participants were asked to rate the extent to which words from Set-1 were related to trust or distrust, from the perspective of trust in general, trust between people, or trust in automated systems, for a total of six between-subject conditions. Participants rated the relatedness of the word to trust or distrust using a seven point scale, with end points of "positively related to trust (or distrust)" and "negatively related to trust (or distrust)."

Results. Participants ratings were analyzed in two ways. First, for each word, average ratings of trust were correlated with average ratings of distrust, for each of the three conditions (general trust, human-human trust, and human-machine trust). As seen in Figures 3.1, 3.2, and 3.3, ratings of trust were highly negatively correlated with ratings of distrust ($r = -.96$, $r = -.95$, $r = -.95$, respectively). Thus, words that had a high positive rating for trust also had a high negative rating for distrust. This indicates that concepts of trust and distrust are in fact opposites, rather than comprising different factors. If any other factors are present, they can explain a maximum of 10% ($1 - 0.95^2$) of the variance in trust ratings. Additionally, we compared ratings of individual words across the three conditions of general, human-human, and human-machine trust, to see how individual words might be differently related to the three types of trust. Words were assigned, according to their average ratings, into the top 5, 10, 15, 20, 25, and 30 words most related to trust and distrust, for each condition. For example, the five words most related to general trust were *honor*, *trustworthy*, *honesty*, *integrity*, and *love*. The five words most related to trust between humans and automated systems were *trustworthy*, *reliability*, *loyalty*, *honor*, and *confidence*. The five words most related to trust between people were *honor*, *loyalty*, *trustworthy*, *honesty*, and *faith*. The degree to which these sets overlap gives an indication of the extent to which concepts of trust and distrust were similar for the three conditions.

Ratings of Unlabeled Trust vs. Distrust, for 112 Words

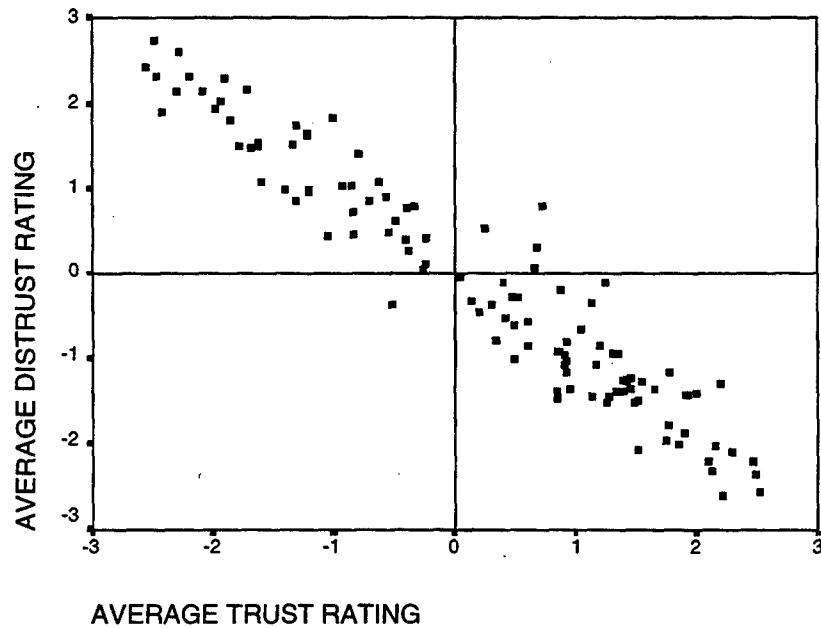


Figure 3.1 Ratings of unlabeled trust vs. distrust, for 112 words.

Ratings of Human-Human Trust vs. Distrust, for 112 Words

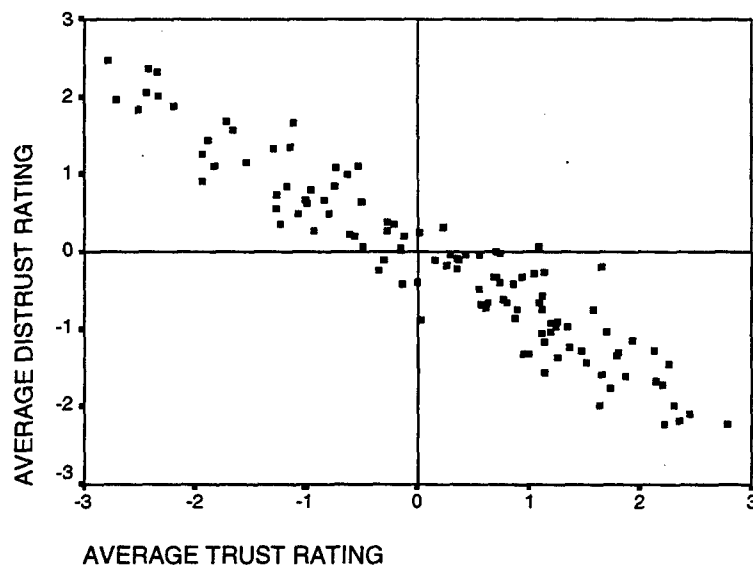


Figure 3.2 Ratings of Human-human trust vs. distrust, for 112 words.

Ratings of Human-Machine Trust vs. Distrust, for 112 Words

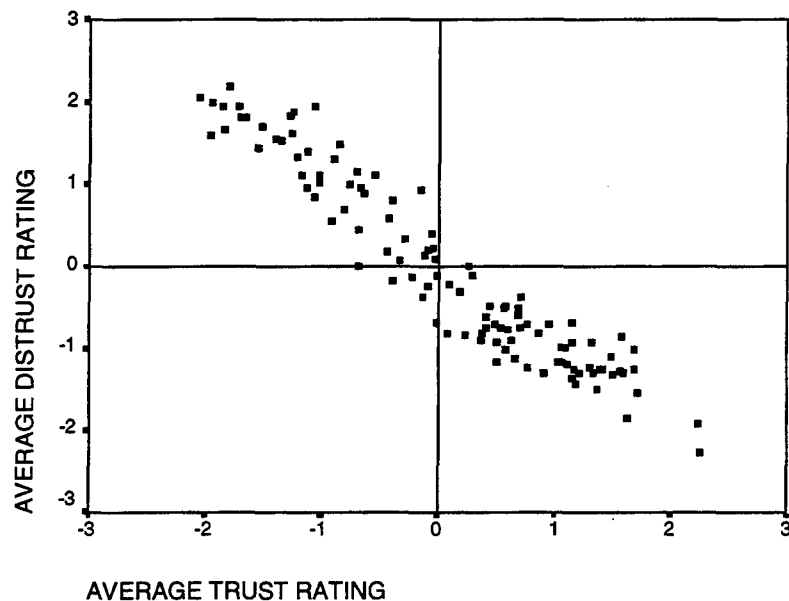




Figure 3.3 Ratings of Human-machine trust vs. distrust, for 112 words.

One measure of this overlap is the size of the union of the sets across the three conditions. For example, if the "top 5" sets for each condition were identical, then the union set size would be 5, indicating the highest possible similarity. If the "top 5" sets were completely different, the union set size would be 15, indicating no similarity across groups. For the "top 5" set then, the minimum union set size would be 5, while the maximum union set size would be 15.

Table 3.6 shows these union sets for the top 5, 10, 15, 20, 25, and 30 words most related to trust (the right half of the diagram) and least related to trust (the left half of the diagram). Note that for 10 of 12 union sets, the size of the union set is 150% or less than the *minimum* union set size. Nine of 12 sets have a union set size that is 50% or less than the *maximum* union set size. These percentages, as well as the union set size and maximum and minimum set sizes, are plotted in Figures 3.4 and 3.5, for the sets of words most negatively and positively related to trust. Thus, while the word sets are not identical across conditions, the relatively small set size compared to the maximum union set size indicates a reasonable degree of similarity across conditions. It should be noted that for the larger word sets, it is more likely that the sets will overlap. Since there were fewer than 90 words that were positively related to trust in the set participants were asked to rate, the sets of 30 words most related to trust had to overlap across the three conditions. However, the degree of overlap was similar across both the small and the large sets.

Table 3.6 Word Sets Related to Human, Human-Machine, and General Trust

 Less Similar to Trust
  More Similar to Trust

Original Set Sizes	5	10	15	20	25	30	30	30	25	20	15	10	5
Trust Rating Range	-2.8...-1.8	-1.8...-1.5	-1.5...-1.3	-1.3...-1.1	-1.1...-.09	-.09...-.07	1.0...1.1	1.1...1.2	1.2...1.3	1.3...1.5	1.5...1.7	1.7...2.5	
Size of Union	7	15	18	26	30	34	39	34	36	23	15	7	
Percent of Min Union	140%	150%	120%	130%	120%	113%	130%	136%	180%	153%	150%	140%	
Percent of Max Union	47%	50%	40%	43%	40%	37%	43%	45%	60%	51%	50%	53%	
Words in Union Set from all 3 conditions (general trust, human-human trust, human-machine trust)	Betray Cheat Deception Distrust Mistrust Phony Steal	Betray Beware Cheat Cruel Deception Distrust Falsity Harm Lie Misleading Mistrust Phony Sneaky Steal Suspicion	Betray Beware Cheat Cruel Deception Distrust Falsity Harm Lie Misleading Mistrust Phony Selfish Skepticism Sneaky Steal Suspicion Wariness	Anger Attack Betray Beware Biased Cheat Cruel Deception Denial Distrust Error Falsity Harm Lie Misleading Mistrust Overcharge Phony Selfish Skepticism Sneaky Steal Suspicion Wariness Wrong	Anger Attack Betray Beware Biased Caution Cheat Cruel Deception Denial Distrust Error Failure Falsity Harm Hesitation Lie Loss Mistake Misleading Mistrust Overcharge Phony Selfish Skepticism Sneaky Steal Suspicion Wariness Wrong	Ambiguity Anger Apprehensive Attack Betray Beware Biased Caution Cheat Cruel Deception Denial Dispute Distrust Doubt Error Failure Falsity Harm Hesitation Lie Loss Mistake Misleading Mistrust Overcharge Phony Selfish Skepticism Sneaky Steal Suspicion Wariness Wrong	Absolute Assurance Certainty Closeness Commit Competence Confidence Cooperation Credit Definite Entrust Faith Familiarity Fidelity Friendship Guardianship Honesty Integrity Intimacy Love Loyalty Moral Pledge Positive Promise Respect Responsibility Security Sincere Stable Surety Trustworthy Understand-ability Unerring	Assurance Certainty Closeness Commit Competence Confidence Cooperation Definite Entrust Faith Familiarity Fidelity Friendship Guardianship Honesty Integrity Intimacy Love Loyalty Moral Pledge Positive Promise Respect Responsibility Security Sincere Stable Surety Trustworthy Understand-ability	Assurance Certainty Confidence Entrust Faith Familiarity Fidelity Friendship Guardianship Honesty Integrity Intimacy Love Loyalty Moral Pledge Positive Promise Respect Responsibility Security Sincere Trustworthy Understand-ability	Assurance Certainty Confidence Entrust Faith Familiarity Fidelity Friendship Guardianship Honesty Integrity Intimacy Love Loyalty Moral Promise Reliability Respect Responsibility Security Sincere Trustworthy Understand-ability	Assurance Confidence Entrust Familiarity Fidelity Friendship Honesty Integrity Love Loyalty Moral Promise Reliability Respect Responsibility Security Sincere Trustworthy Understand-ability	Familiarity Honesty Honor Integrity Loyalty Reliability Trustworthy	

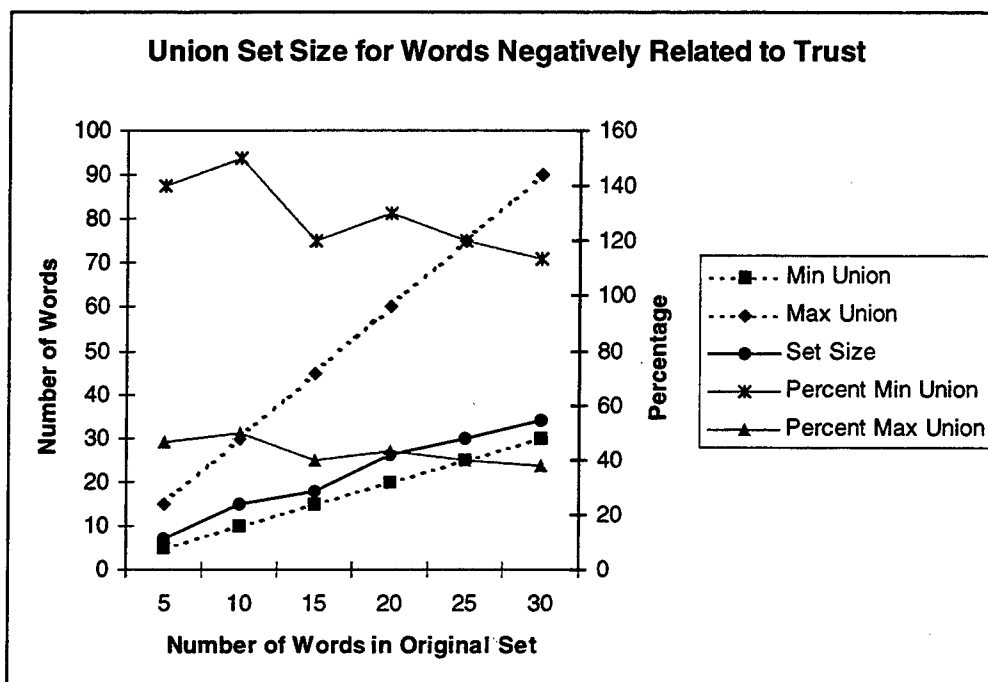


Figure 3.4 Union set size for words negatively related to trust.

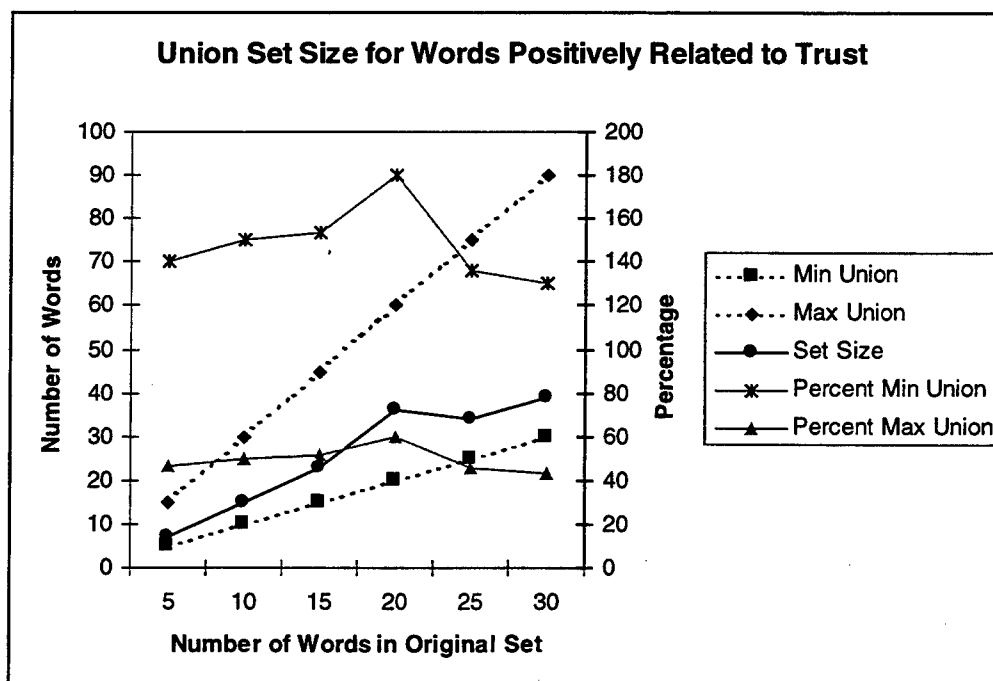


Figure 3.5 Union set size for words positively related to trust.

The above experiments provide results which are important to the development of an empirically developed measure of trust. First, the high negative correlations of ratings of trust and distrust indicate that these concepts can be treated as opposites, lying along a single

dimension of trust. In previous studies, this has been assumed, but not empirically tested. In practical terms, this implies that it is not necessary to develop questionnaires to measure high and low levels of distrust, as well as high and low levels of trust. Second, patterns of ratings were similar across the conditions of general trust, human-human trust, and human-machine trust, as indicated by the high degree of similarity in sets of words related to trust. This indicates that future work on the development of trust measures might not have to treat these types of trust differently, and that results from studies of human-human trust (e.g., those that examine stages in the development of trust; Rempel et al., 1985) may indeed have applicability to situations of trust between humans and automated systems. Again, this transfer of trust concepts from the sociological to human-machine domain had not previously been empirically tested.

Future work (Phase 3) has been initiated to develop a multi-dimensional scale of trust based on the results of these two phases. Participants are being asked to rate the similarity between all pairs of words that were highly positively and highly negatively related to trust, from the results of Phase 2. The results from this paired comparison study will be analyzed using factor and cluster analysis techniques to identify a set of factors which comprise trust. A multi-dimensional scale of trust will then be developed based on these factors. We plan to use this final scale, possibly coupled to performance and/or process measures (see next section) as a measurement basis in human-in-the-loop experiments we hope to conduct in continued research.

3.5 Performance and Process Measures

In addition to subjective measures such as the rating scales described above, one can also investigate trust by considering performance and process measures as measured on operators or decision-makers while carrying out their tasks. Intuitively, if people hold an inappropriate level of trust in a situation, combined human-system performance may suffer. Additionally, as people's trust in various system aspects changes, they may act in qualitatively different ways. For example, if an operator loses trust in the performance of an automated controller, he may make more use of manual control. Similarly, if an operator distrusts some source of information, he may switch to a new information source, or seek to verify the information.

Several researchers have employed performance and process measures in experiments in which system components were manipulated in order to affect participants' trust in the system.

For example, Knapp and Vardaman (1991) performed a study in which reaction time was used as a measure of complacency. The authors suggested that reaction times to warnings would be longer in situations in which they had grown complacent due to possible automated (i.e., system) responses to the warnings. Singh, Molloy, and Parasuraman (1993b) also studied automation-induced complacency by introducing failures in an automated system monitor. Failures occurred when the automated monitor did not detect system malfunctions. Participants were required to identify when the automated monitor failed to detect malfunctions, by detecting those malfunctions themselves. Singh et al. (1993b) found that participants grew more complacent, identifying fewer malfunctions, when the automation performed at a consistent (but not perfect) level of reliability, than when its level of reliability was variable.

The studies by Lee and Moray (1992, 1994) and Muir and Moray (1996) discussed in the previous section also employed process and performance measures to characterize participants' trust of aspects of a simulated, semi-automated process control plant. In these studies, participants had the option of using either manual or automated controllers to control certain process aspects. Measures which were intended to capture aspects of participants' trust in the automated controller included the amount of time spent using the automated vs. manual controller and the number of manual control actions, the overall use of automated sub-systems, and the number of actions taken to monitor the actions of the automated process. It is important to consider the particular situation when applying these measures. For instance, in one experiment (Lee & Moray, 1992) found that use of automated controllers actually increased when subjectively rated trust declined. However, in this case the *system* rather than automated controller performance was degraded, resulting in reduced trust in the *system*. Automatic or manual control actions had the same effects, and participants resorted to automatic control to try to recover from the system fault.

There are particular types of process and performance criteria that are of interest in the IW/AADM domain that are appropriate for exploration in the next research phase. It is particularly important to determine possible cues and indicators that indicate a shift from a state of trust to distrust. One could label these "Indications and Warnings," or I&W criteria. These are indicators that co-occur with the incipient shift towards distrust. If these criteria can be identified, then it is possible that countermeasures (either technical or procedural) could be developed which prevents such shifts, in circumstances where they are inappropriate. An obvious metric, the cessation of use of an automated decision aid, would indicate when a total collapse of trust in the aid has occurred. This is the point where the user has effectively switched off the decision aid. Finally, we also plan to explore the notion of temporal trajectories, or dynamic patterns of trust over time, in order to use those trajectories to provide further insight into trust related behavior, and hopefully develop the means to prevent undesirable human-system behavior. For example, it can be important to minimize the duration of the inertia or hysteresis loops in trust dynamics indicated in previous research. In terms of IW, for a system that is sound, it is desirable to identify techniques to *avoid* distrust, *deter* degradation to a fully manual mode, and to *minimize* the duration of distrustful states.

3.6 Summary

In summary, both subjective rating scale measures, and performance and process measures have been used to measure trust. Trust has been studied in circumstances of interpersonal relationships and with respect to trust in automated systems. However, no one definition of trust has been used consistently across these studies. In fact, it is most appropriate to consider the concept of trust to be multi-faceted, encompassing many different qualities rather than a single-dimension concept. We described the initial findings from a study designed to develop an empirically based, multi-dimensional scale of trust, based on the multiple qualities of trust expressed in prior research. We plan to use this scale in combination with appropriate performance and process measures, to assess people's trust in AADM environments. An experimental framework developed to inform experimentation on these issues is described in Section 4.

3.7 References

- Knapp, R. K. & Vardaman, J. J. (1991). Response to an automated function cue: An operational measure of complacency. *Proceedings of the Human Factors Society 35th Annual Meeting*, 112-115.
- Larzelere, R. E. & Huston, T. L. (1980). The dyadic trust scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and the Family*, 42(3), 595-604.
- Lee, J. D. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10) 1243-1270.
- Lee, J. D. & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153-184.
- Lerch, F. J. & Prietula, M. J. (1989). How do we trust machine advice? In Salvendy, G. and Smith, M.J. (Eds.), *Designing and using human-computer interface and knowledge based systems*. Amsterdam: Elsevier Science Publishers.
- Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automatic systems. *Ergonomics*, 37(11), 1905-1922.
- Muir, B. M. & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429-460.
- Rempel, J. K. & Holmes, J. G. (1986, February). How do I trust thee? *Psychology Today*, 28-34.
- Rempel, J. K., Holmes, J. G. & Zanna, M. P. (1985). Trust in Close Relationships. *Journal of Personality and Social Psychology*, 49(1), 95-112.
- Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35, 651-665.
- Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist*, 26,(5) 443-452.
- Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, 35(1), 1-7.

- Schlenker, B. R., Helm, B. & Tedeschi, J. T. (1973). The effects of personality and situational variables on behavioral trust. *Journal of Personality and Social Psychology*, 25(3), 419-427.
- Sheridan, T. B. (1988). Trustworthiness of command and control systems. *IFAC Man-Machine Systems*, 427-431.
- Singh, I. L., Molloy, R. & Parasuraman, R. (1993a). Individual differences in monitoring failures of automation. *The Journal of General Psychology*, 120(3), 357-373.
- Singh, I. L., Molloy, R. & Parasuraman, R. (1993b). Automation-induced "complacency": development of the complacency-potential rating scale. *The International Journal of Aviation Psychology*, 3(2), 111-122.
- Zhang, L., Helander, M. G. & Drury, C. G. (1996). Identifying factors of comfort and discomfort in sitting. *Human Factors*, 38(3), 337 - 389.

4.0 INVESTIGATING TRUST IN AN IW DOMAIN

4.1 Introduction

Given the tools for measuring trust that were described in the previous section, it is possible to consider the types of empirical studies which could be performed to investigate human trust in IW domains. Such controlled investigations would provide a better understanding of what situation characteristics influence human trust, as measured by either or both of the subjective rating and performance/process measures described above, and also how changes in an operator's trust in system components affects ultimate system performance. In order to develop possible scenarios for investigating aspects of trust in an IW situation, it is instructive to consider how trust has been investigated in other complex, dynamic systems.

4.2 Previous Investigations of Trust in Automated Support Systems

As described in Sections 2 and 3, empirical work in the area of human trust in automated support (decision-aided) systems is limited, and has concentrated primarily on investigating trust in simulated, semi-automated process control environments. Additionally, and significantly, due to our interest in IW environments, these studies have been in *non-adversarial* domains.

As discussed in Section 2, Muir and Moray (1996) and Lee and Moray (1994) studied issues of human trust in simulated, semi-automated pasteurization plants. In these experiments, participants were asked to control a simulated pasteurization process either by controlling pump and heating sub-systems, or by activating an automated controller, in order to produce pasteurized liquid. Different system aspects were altered to see how participants' trust in systems components, such as the automated controller, was affected. In particular, Muir and Moray (1996) altered the quality of the pump systems by introducing either *random or constant errors* in its ability to maintain a set-point, introduced errors into the pump's *display* of its pump rate (although the actual pump rate was error-free), and the performance of the automated controller in setting and maintaining appropriate settings for the pump. Lee and Moray introduced *faults into pump performance* (Lee & Moray, 1992) or *faults into either automatic or manual controllers* (Lee & Moray, 1994). These conditions are not unlike the type conditions that may arise in IW environments—*anomalies* could be introduced at various points in an AADM system, including sensor data or processing algorithms. Trust was measured both subjectively, using rating scales (which were not extensively developed), and objectively, by logging participants' actions (e.g., hypothesizing that more or less use of an automated control system implied more or less trust in that automated system). Because faults were introduced into different components, these experiments investigated trust in a particular system aspect (e.g., the quality of the automation, or the quality of the underlying pump system) rather than trust in automation generally.

4.3 Designing Experimental Scenarios for Studies of Trust in Aided Adversarial Decision Making (AADM) Environments

4.3.1 Developing a Framework for Experimentation

In the Aided Adversarial Decision Making (AADM) environments of interest to this project, DF techniques are used to aid the decision maker by synthesizing data from numerous sources into a form useful for the decision maker. Because the environments of interest are ones involving adversaries, the possibility of corruption in either or all of the data, fusion algorithms, and displays involved in such decision-aiding systems can be introduced by "Information Operations" (Offensive IW operations) carried out by the hostile forces. We propose to conduct human-in-the-loop experiments to study various hypotheses related to human trust under IW conditions and in AADM environments, as affected by such IW environments. In order to guide the development of such experiments, we developed a framework which integrates and systematically varies the various factors which could influence human trust in AADM environments. These factors are drawn in part from an examination of some of the experimental studies cited above.

The multi-faceted manner in which trust was investigated in the experiments described above suggests two dimensions along which studies of human trust in complex environments, such as an aided adversarial decision making environment, could vary: we called these the *system* dimension, and the *surface-depth* dimension.

4.3.1.1 System Dimension:

In the pasteurization experiments (Lee & Moray 1994; Muir & Moray, 1996) the quality of system performance was manipulated at what could be called different "system" levels. Faults or random errors were introduced at the physical environment level - the process control system itself (i.e., the pumps), and at the level of a (system) control device or system - the automated controller. Additionally, faults were introduced into the interface itself. There are analogous levels in an AADM environment. The physical environment level in the pasteurization experiments—the pumps and heaters—corresponds to the actual tactical situation that is taking place. Just as the states of pumps and heaters can be observed and controlled, the states (e.g., current locations, available weapons) of hostile and friendly assets can be assessed, and actions related to the situation can be taken. The next level, DF systems and algorithms, which automatically combine and synthesize information obtained from the tactical environment (forming the basis for control (or decision) actions), can be considered analogous to the automated controller in the pasteurization experiments, which used information from the physical control system to automatically take control actions. Finally, in an AADM environment, one can consider a third level, the interface level. At this level, the results of the DF algorithms are displayed to the operator, in order to aid decision making.

4.3.1.2 Surface-Depth Dimension:

Another dimension along which investigations of trust can vary is a "surface-depth dimension." The surface level corresponds to the information available about the environment

(as formalized in Brunswik's Lens Model; Cooksey, 1996), whereas the depth level corresponds to the actual state of the environment. The manipulations performed by Muir and Moray (1996) can be described in terms of these dimensions. Muir and Moray (1996) manipulated both the characteristics of the pump itself (the depth level) and the display of the pump rate (the surface level). This surface-depth dimension can be applied at all three of the system-dimension levels described above, resulting in six combinations, as shown in Table 4.1.

Table 4.1 Components of an Aided Adversarial Decision Making Environment Described Along a System and a Surface-Depth Dimension.

	Surface-Depth Dimension	
System Dimension	Surface Level	Depth Level
Environment Level: Tactical Situation	Sensed and Observed Data	Evolving Tactical Situation
Intervention Level: DF Algorithms	Results of Algorithms	DF Algorithms
Interface Level: Decision Aid	Display Format	Information to be Displayed

We propose that the combination of the Surface-Depth dimension and three system levels will provide a useful framework for organizing future experimentation in the area of human trust in AADM environments. That is, these Dimension-Level (DL) pairs form a set of experimental factors which could be varied in AADM-IW related experiments. For instance, at the level of the tactical situation (the environment level), the depth dimension corresponds to the actual states and activities of the various players in an evolving tactical situation. In turn, the degree or nature of this tactical-depth factor could be varied in an AADM-IW sense over levels such as "Benign" or "Threatening" or "Critical." The surface level in this case corresponds to sensed or observed information about the environment (i.e., the tactical situation). Again, in terms of AADM, the levels of this tactical-surface factor could be varied over levels such as "Uncorrupted" or "Moderately Corrupted" or "Severely Corrupted." At the level of the DF algorithms (the intervention level), the depth level of the surface-depth dimension corresponds to the structure of actual algorithms and procedures themselves. The surface level reflects the estimates produced by these algorithms. Finally, at the interface level, the depth dimension corresponds to the actual information or advice that is to be given to the operator (the "state" of the display), while the surface dimension, analogous to the DF algorithm case, corresponds to the manner or format in which it is displayed.

4.3.1.3 Further Categories of Corruption:

Within each "Factor" cell of Table 4.1, it is possible to identify various types of malfunction, or causes of information degradation or corruption:

1. Element Degradation. The *quality* of the system component can be degraded through constant, random errors, or discrete failures.
2. Element Failure. System components can *fail* completely resulting in a loss of data.

Different causal factors for the corrupting processes can also be considered:

1. Non-intentional. System components can degrade due to non-intentional malfunction (e.g., material failures, maintenance-related faults).
2. Sabotage. An enemy can take *intentional action* to interfere with a system component or data stream.
3. Subterfuge. An enemy can take intentional action both to interfere with a system component, and to disguise that sabotage.

Given a particular experimental context, the surface-depth and system dimensions described above, along with the two levels of malfunction, and three causal factors, can be used to systematically define a series of experimental manipulations which can be used to investigate issues of human trust in aided adversarial decision making environments. As discussed above, this provides a framework for AADM-IW experimental planning and design.

4.3.2 Experimental Context

For studies of aided adversarial decision making, a possible experimental context would be an interactive battle simulation in which people must make interpretations and/or decisions (e.g., identification of unknowns, decisions to engage hostile forces) based on information gathered and fused into decision-aiding estimates about the situation, such as related to electronic emissions, weapons profiles, and locations and movements of various agents. The simulation would include DF modules which could synthesize environmental information in order to aid the participants decisions.

4.3.2.1 Experimental Scenarios and Manipulations:

The system and surface-depth dimensions, along with the levels of malfunction and causal factors, can be used to identify possible experimental manipulations in the study of trust in aided adversarial decision making, as shown in Table 4.2.

Scenario 1 (Environment/Surface/Degradation/Sabotage). Manipulate the quality of sensed or observed situational data that is being input to the DF algorithms, to simulate adversarial interference in data gathering mechanisms. This would include viral attacks on local/sensor based information processes (e.g., detection processing), emplacement of hostile chips in sensor-related hardware, etc.

Scenario 2 (Environment/Depth/Degradation/Sabotage). Simulate hostile forces' use of deceptive tactics, use of decoys, camouflage (camouflage, concealment, and deception [CCD]), etc. to disguise their true intent.

Scenario 3 (Intervention/Surface/Degradation/Sabotage). Introduce random error into DF processing results to simulate possible attack on these systems by sabotage, computer viruses.

Table 4.2 Potential Experimental Scenarios and Manipulations, Organized by System and Surface-Depth Dimensions and Levels of Malfunction, Causal Factors.

System Dimension	Causal Factors	Surface/Depth Dimension	Surface Level		Depth Level	
		Malfunction Level	Degradation	Failure	Degradation	Failure
Environment	Non-intentional					
	Sabotage		Introduce noise into sensed, observed situational variables (e.g., aircraft flight profile data)		Have hostile forces engage in deceptive tactics	
	Subterfuge					
Intervention	Non-intentional					
	Sabotage		Corrupt the information being passed from the algorithms to the decision aid		Simulate viruses in fusion algorithms by introducing random error, incorrect rules into algorithms	
	Subterfuge					
Interface	Non-intentional					
	Sabotage		Vary fidelity of display to indicate "correct" level of trust in information		Introduce corruption (e.g., random error, faulty advice) in information provided by decision aid	
	Subterfuge					

Scenario 4 (Intervention/Depth/Degradation/Sabotage). Exploit knowledge of DF algorithms and knowledge-based system structures to cause faulty behavior, such as the insertion of biases, correlated data, non-normal data, or incorrect rules into the knowledge-based systems. Manipulate the quality of processed information that is being input to the decision aids, in order to simulate adversarial interference in these data transfer mechanisms.

Scenario 5 (Display/Surface/Degradation/Sabotage). Alter the properties of the decision aid display in accordance with the level of trust the adversary desires the friendly agent to have in the

information being provided.

Scenario 6 (Display/Depth/Degradation/Sabotage). Introduce errors in the decision aid's reasoning mechanisms to simulate possible attack on these systems by sabotage, computer viruses.

4.3.2.2 Experimental Participants:

Within these possible scenarios, experiments could be constructed with either single or multiple participants. In a single participant scenario, participants would interact with a simulated tactical situation, obtaining information and taking actions. Adversarial IW activities, such as those indicated in the above scenarios, would occur automatically through the simulation, based on a pre-defined script.

In a multiple participant scenario, there are several possibilities. Two or more participants could collaborate against simulated hostile forces and IW attacks, as in the single participant case. However, with multiple players, it would be possible to investigate how people integrate information they obtain through their own systems, and information they obtain "second-hand" from other people, and how trust in that information may be differentially affected by its source (or "pedigree"). Alternatively, participants could compete against each other, with IW attacks introduced either automatically through the simulation, or at the discretion of the participants.

In either of these cases, it is also possible to consider that some agents or participants would be synthetic; that is, created in software as so-called "intelligent agents." This would add a dimension of experimental control, since the behavior of the agent would be fully controllable, or at least controllable within known limits.

4.4 References

- Cooksey, R. W. (1996). *Judgment analysis: Theory, methods, and applications*. San Diego: Academic Press.
- Lee, J. D. & Moray, N. (1992). Trust, control strategies, and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243 - 1270.
- Lee, J. D. & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153 - 184.
- Muir, B. M. & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429 - 460.

5.0 IMPLICATIONS FOR THE DESIGN OF AN INFORMATION WARFARE LABORATORY

5.1 Overview

In our previous chapters we have developed an overview of the state of research in the area of human trust in automated systems, with a particular focus on the implications of trust and distrust on decision making in adversarial and IW environments. As described in the previous sections, there are two primary bodies of work that are potentially related to trust in AADM and IW environments: 1) Trust studies, carried out by sociologists and human factors engineers, which focus on interpersonal trust, and on trust in automated systems in non-adversarial environments, respectively, and 2) studies of IW effects on AADM, which have generally focused on Defensive IW (information protection), taken a top-level point of view on such matters as policy, concepts, impacts, and concerns, as well as considered human decision making in AADM environments using Boyd's "OODA" loop model ("Observe-Orient-Decide-Act"; 1987). Figure 5.1 shows this state of research diagrammatically, showing that to our knowledge little to no experimental work has occurred that addresses both trust and the AADM domain.

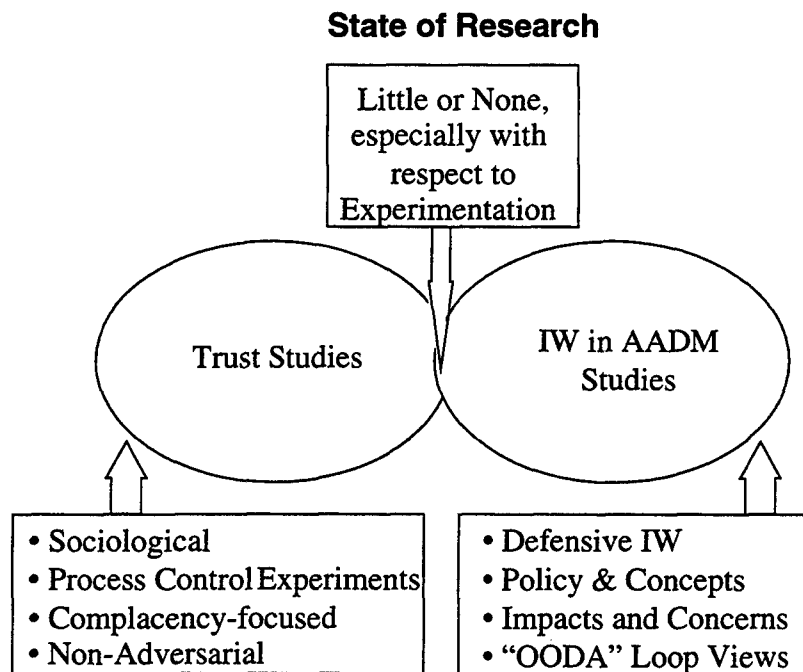


Figure 5.1 Current state of research on trust and IW.

From the investigations on trust, we concluded that trust is best seen as a multi-dimensional construct, reflecting a set of interrelated perceptions (e.g., the reliability, or predictability of an entity) and actions (e.g., use of an automated system, reliance on a person). The concept of trust is based on past experience of a person with the entity to be trusted, characteristics of the entity (e.g., is it predictable, are its mechanisms understandable), and characteristics of the person (e.g., in the case of automated systems, someone might "overtrust"

the automation if he lacks the skills to take over manually). Trust also has dynamic characteristics, changing over time, and in response to specific events. It can increase over time as people have experience with a reliable and predictable system, and then degrade if that system exhibits faulty behavior. Once trust has degraded, it can take time before trust in the system can be regained. Trust as a construct is important in AADM environments insofar as it impacts changes over time in the behavior of a decision-maker, with respect to their use of information and decision aids.

However, there have been no experimental studies of trust which have specifically addressed trust in adversarial environments, and nor have the studies of IW in AADM addressed issues of trust. We believe that people's reliance on information and automated algorithms, and in particular, the event- and time-driven patterns of trust development and diminishment may vary significantly between adversarial and non-adversarial situations.

5.2 Laboratory Design for Pilot Studies

In order to empirically investigate issues of human trust in AADM situations, it is necessary to construct an experimental test bed. This test bed, a dynamic interactive computer micro-world, should comprise a battlefield and DF simulation, configurability, display and controls, data logging features, and experimental scenarios, as described below.

Battlefield and Data Fusion Simulation. A discrete-event simulation of a battlefield environment, containing multiple, dynamic, and possible uncertain information sources (e.g., position, electronic emissions, weapons capabilities) about potential threats and friendly assets, along with a decision aid based on data fusion technology is necessary to provide interactive, experimental scenarios to participants. Various experimental scenarios, including dynamic sensor and state variable values, and time and/or action dependent simulated adversarial or non-adversarial events (e.g., information manipulation or degradation) could be developed as a basis for experimentation. Ideally, these testbed components would allow for two-sided simulation between hypothetical friendly and adversarial force commanders or staff, and be "reactive," in that each side's responses to runtime simulation dynamics would reflect actual runtime responsiveness or adaptations—said otherwise, neither side's functions would be "scripted" (non-reactive). In the long run, we would also desire that the testbed incorporate so-called "intelligent agents" or an equivalent mechanism to represent hypothetical or surrogate human test subjects; such capability would permit additional test control, repeatability, and also permit modeling of subject characteristics that might not otherwise be available (e.g., culturally-based features, hostile military doctrinal features).

Configurability. In order to manipulate the independent variables indicated by the experimental framework described above, the experimental system will have to provide the ability to make several kinds of experimental manipulations. The intent here is to allow experiments in which errors or other types of degradation are introduced into the fusion-based decision aid, in order to measure participants' response (in terms of both trust ratings, and observable actions). It must be possible to introduce system events at the different stages of the DF process (e.g., a sensor failure, or a degradation of algorithm output) during the course of a simulation, in order to simulate adversarial and non-adversarial failures in various levels of the

system and decision aid. Also, it must be possible to have access to the inputs and products of the process stages, so ultimately they can be displayed to the decision-maker (participant).

Display and Control. To understand the effects of information manipulation and degradation on performance, it is necessary to have an experimental set-up which allows participants to both perceive the possible manipulations, and change their strategies (e.g., take different actions) based on these perceived abnormalities. Therefore the experimental system must allow participants to see the inputs to and outputs from different stages of the decision aid (i.e., through a display), and to take action based on that information (i.e., have some form of control). More specifically, the system must allow participants to obtain information from different sources and stages of processing, so they can perceive abnormalities in any one source or stage. Additionally, the system must allow multiple paths of action (e.g., different information search strategies, or decisions made with and without the help of the decision aid) to assess how decision making strategies might change based on changes in trust in a particular system component.

Data Capture. To assess experimental performance, it will be necessary to automatically log participants' interactions with the experimental system. Capturing data "within" the simulation system boundaries should be straightforward and would involve seeding the system with software probes for such data capture. These boundaries include the entire simulation software system, and so include means for capturing both parameter values within any software function during a run, and also any interactive actions taken by the operator, measured at the human-computer interface. Additionally, it would be desirable in a robust configuration to have the means to capture physiological parameters for a human subject during any run; this would involve eye-movement, video capture, etc. Finally, we see a need for either within-run or post-run capture of survey/questionnaire type data, related, for example, to capturing observations related to the trust attributes discussed elsewhere in this report.

5.3 A Specific Laboratory Concept

The State University of New York at Buffalo has, through the generosity of Ball Corporation (Dayton Office), received a copy of a simulation system called the "Semi-Automated Ground Environment" or "SAGE." SAGE has the ability to "lay down" both air and ground target-related problem simulations (i.e., scenarios that define the truth of an adversarial behavior at the platform and weapon level). It can also simulate commander types at a coarse level of fidelity. SAGE also has moderate-fidelity simulations of typical military sensor types. Overall, it is a reasonably capable and respected (validated) simulation package. However, we have not yet achieved in-depth knowledge about SAGE since it is a fairly large software system having relatively little and current documentation. Nevertheless, we have achieved a degree of familiarity with it and it may be possible to use it for the experiments characterized in the materials of Sections 3 through 5 herein. Even if it is not the baseline from which we build an experimental environment, considerations about SAGE are typical of what has to be done with any simulation-based approach. (SAGE comes in different configurations, from a full-capability version to a somewhat less capable personal computer version; whether any of these versions will be used is uncertain but there is some flexibility in trading off complexity vs. capability.)

Conceptually, we desire to create a synthetic environment as depicted in Figure 5.2. This simulation concept builds upon our AADM model of Phase One, the figure depicting this model is replicated here as Figure 5.3 to show the similarity among the two. Figure 5.2 shows a full two-sided simulation environment in which friendly and adversary are reactive to each other over time. Each has a decision aid (DA) data from particular sensors, each DA is fed data from particular sensors, and each has a particular display subsystem. SAGE does not incorporate "intelligent agents" but these could be added as an enhancement to achieve more experimental control if desired.

Synthetic AADM Model with SAGE

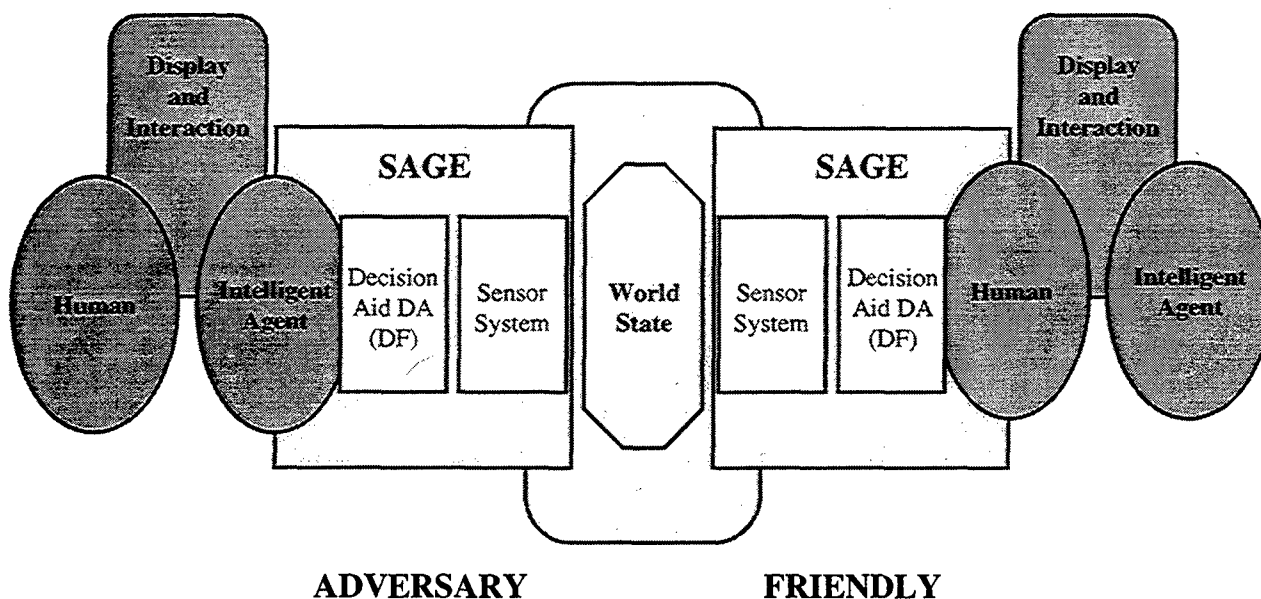


Figure 5.2 Synthetic AADM Model with SAGE

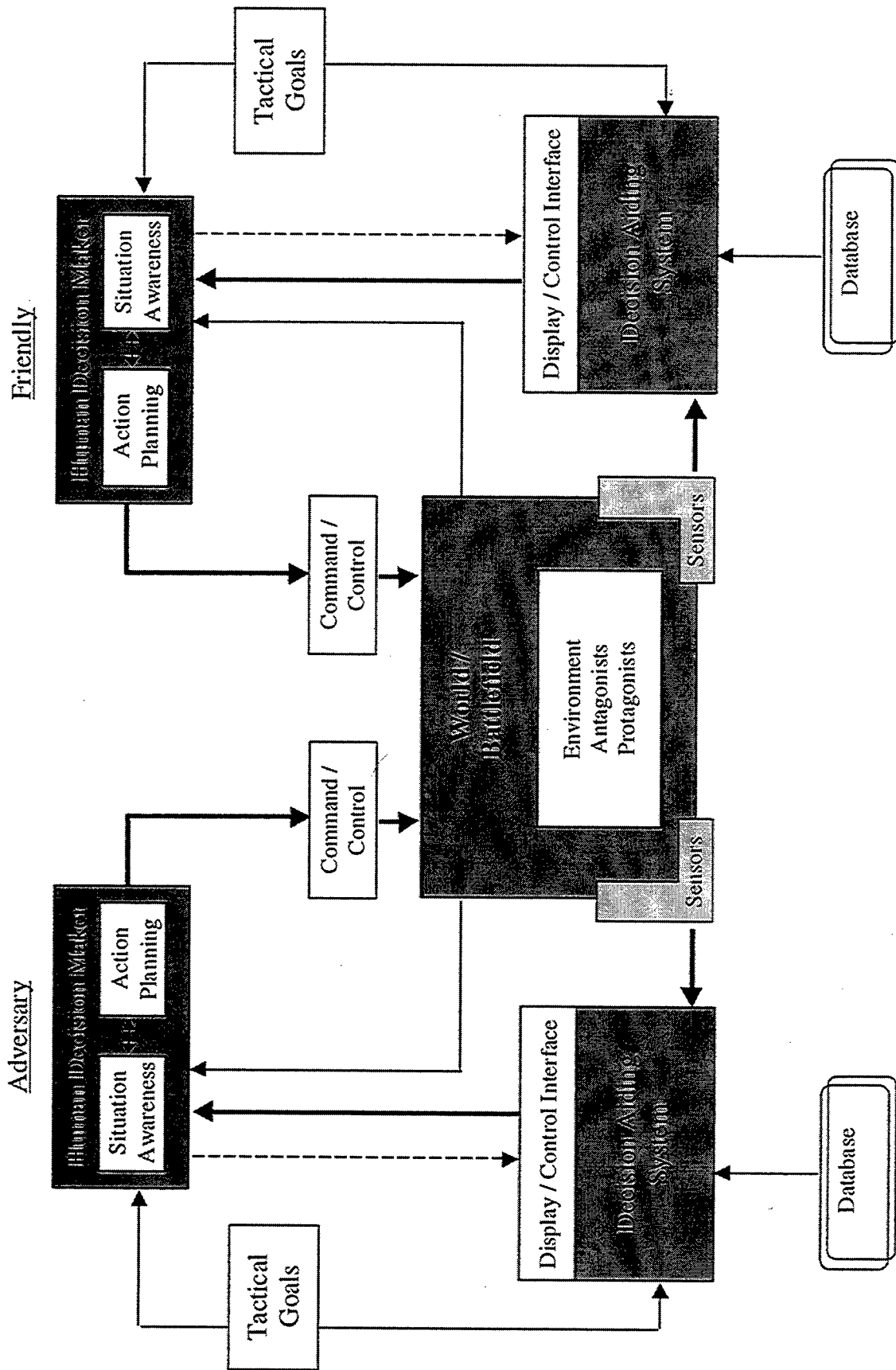


Figure 5.3 Two-Sided General Model

There are a number of other factors of a simulation environment that must be considered in a prototype design among other effects:

1. The capability to simulate Offensive and/or Defensive Information Operations for either side; this is tricky in that it can likely lead to highly-classified aspects of both U.S. and foreign IW capabilities and techniques, but some reasonable representation can likely be carried out in the simulation.
2. The enablement of a means to capture and display the various trust or other metrics decided upon for these prototype experiments; this is usually reflected in an "Analysis Module" or the equivalent in a simulator design; at present, SAGE does not have these trust-related capabilities.
3. The selection of an experimental methodology and approach to experimental designs (e.g., Monte Carlo aspects, statistical experimental design, factors and levels, subject characteristics and blocking factors)

If we consider the "framework" for experimentation described in Section 4, coupled to the simulator design of Figure 5.2, we achieve something like that shown in Figure 5.4 on the following page. This figure shows one side of the two-sided diagram of Figure 5.2, and shows the "entry points" for experimental factors as defined in Section 4. Exactly how to create the IW effects desired will need some further study, but the diagram makes it clear that the overall experimental framework discussed in Section 4 is sound. From a test-capability point of view, the components that would need to be added to the current SAGE baseline are

- simulation of IW operations
- collection of IW metrics at various points in the processing system
- capturing aspects of human-system interaction and monitoring
- ability to selectively generate various types and locations for IW "intrusions" in the overall process
- simulation of "agent" behavior

With limited resources, it is likely that a simpler design for an initial lab prototype will be developed. It is proposed that such prototype be defined in discussions with Air Force staff so that proper priorities can be assigned to the most-desirable simulation features that can be achieved within project resources.

5.4 References

- Boyd, J. (1987, August). *A discourse on winning and losing* (Report No. MU 43947). Maxwell AFB, AL: Air University (unpublished briefing, available through Air University Library).
- Llinas, J., Drury, C., Bialas, W. & Chen, A. (1997). *Studies and analyses of vulnerabilities in aided adversarial decision-making: Final Phase I Report*. Buffalo, NY: SUNY at Buffalo, Center for Multisource Information Fusion.

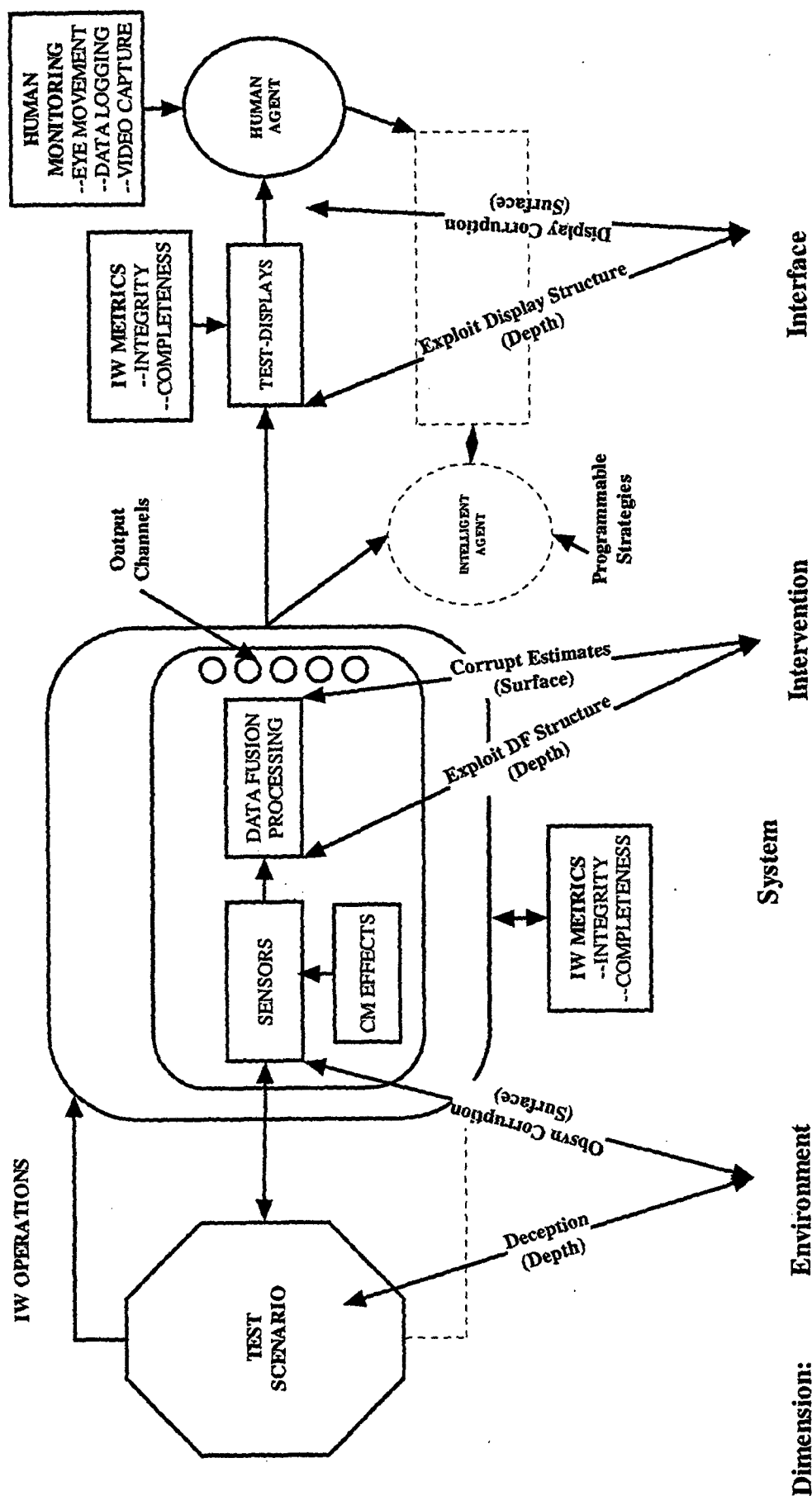


Figure 5.4 Detailed AADM Lab Concept Using "LENS" Model IW Framework

GLOSSARY

AADM	Aided Adversarial Decision Making
ASRS	Aviation Safety Reporting System
C3I	Command, Control, Communication and Information
CCD	Camouflage, Concealment and Deception
CMIF	Center for Multisource Information Fusion
COA	Course of Action
DA	Decision Aid
DF	Data Fusion
DISA	Defense Information Systems Agency
DL	Dimension-Level
DM	Decision Making
FMS	Flight Management System
ID	Identification
IW	Information Warfare
JDL/DFG	Joint Directors of Laboratories Data Fusion Group
MIM	Mixed Initiative Model
OODA	Observe-Orient-Decide-Act
RPD	Recognition-Primed Decision Model
SAGE	Semi-Automated Ground Environment
UB	State University of New York at Buffalo

APPENDIX A

Further Thoughts on Sheridan's Criteria and the IW + AADM Environment

This short appendix contains some further thoughts on how the Trust criteria of Sheridan (1988) can be interpreted and elaborated upon in the context of IW and AADM.

1. One aspect of trust in automation seems to relate to notions of Reliability. That is, a sense of Predictability based on repeatability or consistent functioning. Said otherwise, a sense of reliability in a decision aid is established when it can be observed to create the same output under a particular set of circumstances. If we consider that there is a true input (or "circumstance") I_t , an observable input I_o , and a displayed output O_d , then generally, in the Information Warfare setting, there are various points of vulnerability.
2. I_t is susceptible to deception and corruption of Operations and Actions (false actions in the real world, such as feints)—this is artificially creating a false circumstance.
3. I_o is susceptible to Parametric or Algorithmic corruption—this is creating false output from internal processes in spite of a "correct" circumstance in the real world.
4. O_d is susceptible to Algorithmic corruption—this is creating a false display in spite of both (a) an unperturbed real world and (b) unperturbed (non-display related) processing operations.
5. Another aspect of Trust is a sense of Competence or Robustness in the decision aid (i.e., correct processing in the face of varying—and presumably true—real-world circumstances). In some sense this represents a "calibration" of the decision-aiding software processes/algorithms. The user sometimes sees this through some level of participation in the system development process for the DA—that is, through observation of DA performance during developmental testing. Here, each instance of output would be looked at in the Reliability sense, to the extent that repeatability tests are made. Some users may never see the DA until delivered to the operational setting, in which case they usually see only a relatively small subset of conditions over which the DA is exercised.⁶ The literature about Trust and the related notion of Complacency reflects these remarks with studies that show that unwarranted Trust is often ascribed to DA's on the basis of small-sample conditions.
6. Now the corruption of a users Trust in the DA through IW techniques places a requirement on the deceiver to know the nature of adaptivity in the DA internal workings—he must hypothesize about (or, e.g., rely on intelligence for) these inner workings. But if he is astute in the likely technology for the DA, a reasonable approximation to the inner workings may be feasible. Conceptually, this allows the deceiver to extend the vulnerability discussed above (under Reliability) "n-fold" where n is the range of circumstances for which he is aware of the DA/algorithm workings. Alternately, if the deceiver is aware of the limits of capability of the DA (he could base this on technology assessments as just mentioned), he could create

⁶ Of course, this is dependent on the stochastic variation in real world circumstances; the user may possibly see the DA performing well across a wide variation in circumstances but on average he will see relatively small variations.

circumstances which are in those fringe operating regions and reduce the users impressions of the competence of the DA.

7. Familiarity is yet another feature or aspect of Trust. For automated DA's this can involve the use of common or perhaps military-standard symbology, nomenclature, etc. It can also mean some sense of "naturalness" in operation, in communication, etc. It seems that there may be (at least) two dimensions to Familiarity: a cultural dimension, and a "standards" dimension. The latter would be easy to deceive (e.g., in creating a false display but using military standard symbology) since such matters, even if classified, are usually not highly classified and protected. The former involves much more of an investment on the part of the deceiver—to become culturally "transparent" can require a lot of work. So, it could be that "un-naturalness" in DA workings or displays could be a clue as to possible loss of DA system/information integrity—that is, to the existence of an IW attack.
8. Another dimension is in Understandability. This is particularly important if, as noted in (3) above, the user does not participate in the DA development process (since he will not have been a part of that process and not have had involvement and insight into the innards of the DA). If this is the case, then the degree of Understandability governs the users ability to form his own mental model of the inner workings of the DA, which also lends itself to Predictability. Note that Understandability is not equal to Familiarity, although Familiarity aids in developing an Understanding. Here the deceiver can simply have the goal of generating randomness in his attack, since irregular (hard to understand) patterns between input and output will aid in loss of Trust.
9. On explication of Intention. Estimating Intent is one of the more difficult things to do in military situations but is also, if done correctly, one of the highest payoff areas. Intent is (approximately) an explicit indication of a future planned action to which the actor is committed; the difficulty is in defining and observing those indicators. An actor develops Intent based in part on his value system; that is, he will plan actions that have a sense of payoff against or in the context of a value system. The value system is, however, multi-dimensional—it has relatedness to military goals and objectives (these are the easier elements to estimate) but also to notions of personal value (does this action imperil my hoped-for promotion?), and the societal notions of value we all grow up with (will this action cause harm to anyone?)—these are the harder elements to judge. The development of an Intended action is dependent on a particular or sequence of particular outputs (from a DA) in the context of judged value.
10. Action $(T + \Delta T) \sim [DA\ Output\ T, DA\ Output\ (T-1), \text{etc.}, \text{and Implied/calculated Value of Action, given the DA Output}]$
11. By creating any false output of the DA (by any means), the deceiver corrupts this relationship and leads the deceived toward taking an alternative action. The notion of Value above could also be labeled "Policy." Based on a priori analyses, there could be policies set down which declare that if "This" (an outcome or estimate of a situation—as e.g., from a DA Output, as above), then do "That" (i.e., take Action [x]).

Usefulness is another trait associated with Trust. This relates to the ability of a DA to respond in a "useful" or "responsible" way. This relates also to notions of Value (in the large) but also to notions of Utility of the DA itself, in a local or specific sense. By this latter remark is meant that

if the DA performs and produces output that is consistent with its planned Concept of Employment, then it will be judged "Useful" in this sense (operating as designed). But Usefulness also takes us back to the notions of Value—DAs can be Useful, but not Valuable. The degree to which Usefulness and Value are equivalent depends (again) on the degree to which users have been involved in the specification, design, and development (including testing) of the DA. Most deceivers would probably not have insight into notions of Usefulness, since this can mean many things—but they may have some notions of Value which can be exploited as described above.

We also see from Wickens (1994) that:

Likelihood of use of automation \sim {Trust in auto. / Self-confidence of operator}

That is, we see that the use of automation seems to vary directly with trust and inversely with operator self-confidence, which implies a number of things. For example, if the users are known to be poorly trained then their self-confidence would presumably be low and their reliance on automated DAs high, perhaps to an unwarranted degree. Also, if the DA has only been built to handle easy cases and has never confronted the (rare) fringe-condition—where it can fail—the user may develop an unwarranted Trust in its competence. DAs that are overly automated, to the point of not incorporating much user involvement, can also create situations or patterns of use where user skills atrophy, self-confidence goes down, and (unwarranted) use/Trust goes up.

In Table A.1, we have also summarized Sheridan's trust characteristics from various viewpoints. The first three columns of the table are basically a synopsis of Sheridan's assertions. The last three columns attempt to provide some perspectives on the informational dependencies and vulnerabilities of each trust element, and some ideas on possible metrics which could be used in the proposed human-in-the-loop experiments of a possible next phase. If we summarize some of the potential techniques that an adversary could use in his Information Operations to create distrust, we see that a list such as shown in Table A.2 evolves from reviewing the "Information Vulnerabilities" column of Table A.1. It can be seen that there are two basic pathways to creation of distrust: the Direct IW or "attack on external perception" that can be created by deceptive actions, etc., fully controllable by an adversary—these basically create distrust in the Reliability and Robustness attributes of trust and cause the user to suspect the veracity of the DA across different operating conditions; the other is the Indirect IW attack, which offers the chance for corrupting many more trust attributes but at the expense, to the adversary, of developing the (covertly-gathered) intelligence required to do this, or at least do it well. However, general knowledge about how DF processes work and are typically implemented, can, without system-specific intelligence, allow for a generalized IW attack on the data fusion DA, albeit with some lesser degree of effectiveness.

Table A.1 Notions of Trust and Related IW Features

Trust Attribute	Characteristics	Basis for Trust Development	Information Dependencies	Information Vulnerabilities	Measures and Metrics
Reliability	Repeated, consistent functioning	Conditioning	DA performance depends on "characteristic" or "problem-encoding" parameters (i.e., if SW "reads" problem as "same," given the input data)	Adversary creating false values of problem-characterizing parameters, causing DA to switch problem-modes	<ul style="list-style-type: none"> Bayesian approach (see Sheridan) Reliability can also be linked to Utility: $U = Rel \times Usefulness$
Robustness	Similar to Competence; good performance under varying problem conditions	Conditioning; often achieved with a small sample of cases (and potentially unwarranted)	DA performance depends (more critically than for Reliability) on "characteristic" or "problem-encoding" parameters	Adversary creating false values of problem-characterizing parameters, causing DA to switch problem-modes	Degree of trust as a function of inter-problem performance (or output) variance is one idea; that is, $T = f(Perf\ Var)$
Familiarity	Employment of familiar/friendly/natural procedures, terms, etc.	Inherited from the culture of the domain (or life experience); can be unwarranted, irrational	Familiar notation and symbology	Creation of unfamiliar notation/symbology	Degree of intervention related to adjustments in notation and symbology
Understandability	Not the same as Familiarity or ease of use; supports formation of mental models and predictive framework	Some notion of "visibility" into inner processes. Distrust and alienation formed when inner workings not understood.	Ability to achieve the "visibility" mentioned—availability of information/data related to inner processes	Any actions that lead to bewilderment in viewing inner processes, or reductions in "communicativeness" of the system	Degree of intervention to access inner processes or to reformulate system output
Explication of Intention	Not same as Understandability in that intention is <i>overtly stated</i> ; Understandability allows <i>inferring</i> intent	Execution of actions per stated intent—in a sense, follow-up on intended/stated actions	Stated intent	False indications of intent, or failures in acting as (presumably) intended	Degree of inquiry into intended (but not visible) follow-up actions. Queries related to disparities in inferred vs stated intent
Usefulness	Classical notion of utility	From use of a "focused" or "directed" system that supports assigned work.	Information that supports efficient rate of progress toward work-goal	Information that impedes work-flow	Degree of intervention to alter flow of operations
Dependence	Reliance on system to support human performance	Can be "Catch-22" type relationship—a type of <i>forced</i> trust on a system you (must) depend on	Information that prevents alternative ways of conducting necessary work	Information that suggests that alternative ways of doing work are possible	Number of work-mode changes (e.g., automated to manual possibilities)

Abbreviations: DA=decision aid, SW=software, T=trust or trustworthiness

Table A.2 Possible Information Operation Attacks on Trust Attributes

Trust Attribute	Type of Information Operation to Compromise Attribute
Reliability	Whatever "makes the problem look different"; this could range from Deception to internal attacks on the DF decision aid
Familiarity	Creation of unfamiliar notation or symbology; this would come from internal attacks (i.e., Indirect IW), and would require the adversary to have an idea of the standard (familiar) symbology being used (i.e., a priori intelligence of this DA design feature)
Understandability	Internal attacks on the basic operations of the DF software and in addition its modes of interaction with the user. This too requires the adversary to have a priori intelligence about the DF process design and the operations of the software and its HCI aspects.
Explication of Intention	This is the "Intent" of the data fusion DA, not that of the true adversary. Thus, confounding of this aspect of the user's trust in the DA requires the same action as for confounding Understandability.
Usefulness	Information attacks that impede work flow or create the wrong DA product. These attacks also require some degree of insight into the workings and employment patterns (concept of operations, or ConOps) of the data fusion DA.
Dependence	This feature could allow for creative adversarial actions that lead the user into a dependency (complacency) pattern and then confound the DA at some critical moment. Here, too, some a priori intelligence is required about the planned or designed use patterns of the DA

REFERENCES

- Sheridan, T. B. (1988). Trustworthiness of command and control systems. *IFAC Man-Machine Systems*, 427-431.
- Wickens, C. D. (1994). Designing for situation awareness and trust in automation. *IFAC Integrated Systems Engineering*, 365-370.

APPENDIX B

Summary Literature Review

The following tables have been provided to summarize some of our literature review in succinct format. Basically, there are two bodies of literature summarized here: the Social-Psychology Literature (Table B.1) and the Human Factors Literature (Table B.2). For each, the tables are similarly formatted, and identify or describe:

- the reference
- the Problem Domain or context of the work
- any Definition of Trust provided (explicitly)
- the Basic Model of Trust asserted or used
- the Parameters or attributes asserted or discussed
- any Measurements or Findings of the work

Nine references are reviewed in Table B.1, and eight are reviewed in Table B.2. The references are shown in the reference section.

Table B.1 Social-Psychology Notions of Trust

Literature	Problem domain	Definition	Basic model	Parameters	Measurements/Findings
Deutsch (1958, 1980)	1. Trusting behavior 2. Interpersonal game	Involves the notion of motivational relevance as well as the notion of predictability (or expectation)	1. Social suspicion 2. Mutual trust	1. Variation of the prisoner's dilemma 2. Gains of losses incurred by each person are a function of the choices made by one's self 3. Different 'position' (e.g., a subject played the game twice with the other subject, each time in a different order - position)	1. Two-person non-zero-sum game 2. F scale 1. Motivational orientation, communication, types of power relationships, third parties, and personality are related to trust 2. Subjects tend to be trusting and trustworthy or suspicious and untrustworthy in an ambiguous situation with unknown other 3. Both personality predisposition and internalization of a reciprocal pattern of interrelationships with another are tapped by the game
Gill and Butler (1996)	Trust and distrust in joint-venture	Trust is procedural or impersonal	1. Based on perceived reliability of the individuals to fulfill expectations (an outcome of the nature of social networks) 2. Based on the perceived reliability of formalized systems and computations for making reliable decisions (an outcome of the rules and industrial recipes that pertain in the joint-venture)	An intermediate variable	
Hamsher, Geller, and Rotter (1968)	Interpersonal trust and belief in internal and external	Same as Rotter (1967)	1. Investigate personality variables in the public reaction 2. Trust is a predictor within sexes 3. Internal-external control are only for males	1. Internal control 2. External control of reinforcement 3. Individual differences	1. Personality test 2. Warren Commission Questionnaire 3. Interpersonal trust scale 4. Internal-external scale

Table B.1 Social Psychology Notions of Trust, cont'd.

Literature	Problem domain	Definition	Basic model	Parameters	Measurements/Findings
Larzelere and Huston (1980)	Interpersonal trust specific in intimacy	An integral feature of human relationships; a belief by a person in the integrity of another individual. Trust exists to the extent that a person believes another person (or persons) to be benevolent and honest.		1. Benevolence 2. Honesty	Dyadic Trust scale
Rempel, Holmes, and Zanna (1985), Rempel and Holmes (1986)	Interpersonal trust in close relationship	Trust is seen to evolve out of past experience and prior interaction; feeling of confidences and security in the caring responses of the partner and the strength of the relationship	1. The type of attributions drawn about a partner's motives 2. Hierarchical pattern	1. Predictability 2. Dependability 3. Faith	1. Questionnaire 2. Trust scale Trust is related to personal motivation and is an important way to the success of a close relationship
Rotter (1967, 1971, 1980)	1. Interpersonal trust 2. Test-retest reliability for long periods of time	An expectancy held by an individual or a group that the word, promise, verbal or written statement of another individual or group can be relied upon	1. Marlowe-Crowne Social Desirability scale 2. Lickert-type scale		1. Questionnaires 2. Interpersonal trust scale 3. Two-person non-zero-sum game Trust and trustworthiness showed a significant relationship
Schlenker, Helm, and Tedeschi (1973)	Interpersonal trust	A reliance upon information received from another person about uncertain environmental states and their accompanying outcomes in a risky situation	Interpersonal trust scale (Rotter, 1967)	1. Personality variables 2. Situational variables	2x2 payoff matrix Those who scored high in Interpersonal Trust scale tended to trust the promises of the simulated payer more frequently
Strickland (1958)	Interpersonal trust	One's perceptions and evaluations of the other. The nature of the trust is perceived as 'internal' or 'external' by a man to other person	1. Trust one's subordinate less than he would had the perceived the compliance as stemming from some cause lying within the subordinate himself 2. Make greater efforts to keep him under surveillance	1. Person's tendency to trust another 2. Motivation	1. Questionnaires 2. Payoff matrix 3. A man can't know firsthand the nature of the loyalty of the others until the man perceives that the others have opportunities to be disloyal

Table B.1 Social Psychology Notions of Trust, cont'd.

Literature	Problem domain	Definition	Basic model	Parameters	Measurements/Findings
Wrightsmann (1966)	Interpersonal trust		<ol style="list-style-type: none"> 1. Detect some of the attitudinal and personality correlates of trusting and trustworthy behaviors in an interpersonal situation 2. Determine whether the expectation of winning real money influences the frequency of trusting and trustworthy behavior in a two-person non-zero-sum game 	Real money instead of imaginary money	<ol style="list-style-type: none"> 1. Two-person non-zero-sum game 2. Several sociological trust scales <ol style="list-style-type: none"> 1. Real/Imaginary money did not influence the frequency of trusting behavior 2. The size of reward has little effect on subject's behavior 3. Attitude and personality might relate to game behavior

Table B.2 Human Factors Notions of Trust

Literature	Problem domain	Definition	Basic model	Parameters	Measurements/Findings
Bliss (1997)	<ol style="list-style-type: none"> 1. Mistrust 2. Cry-wolf effect 		Signal Detection Theory	<ol style="list-style-type: none"> 1. Reaction mode: vocal or manual 2. Reaction type: alarm response or cancellation 	<ol style="list-style-type: none"> 1. Complacency reaction time 2. Voice-activated reaction time is faster than manual reaction 3. Participants responded to a greater number of alarms when using voice than when pressing key
Chambers and Nagel (1985)	<ol style="list-style-type: none"> 1. Human error 2. Information transfer problem 		<ol style="list-style-type: none"> 1. Whether the pilot is a perpetrator or an agent? 2. Whether the pilot is fundamentally at fault or merely the last step in a sequence of events beginning with the design of the aircraft and its operational system 		

Table B.2 Human Factors Notions of Trust, cont'd.

Literature	Problem domain	Definition	Basic model	Parameters	Measurements/Findings
Knapp and Vardaman (1991)	1. Cockpit automation 2. An operational measure of complacency		Primary/Secondary tasks	Failure warnings	1. Flight deck simulation 2. Complacency reaction time
Lerch and Prietula (1989)	1. Confidence Rating 2. Effects of source pedigree of problem solving advice on self-reported measures of agreement with the advice	Rotter (1980), Rempel et al. (1985), Muir (1987)	Hierarchical pattern in terms of the level of attributional abstraction demands, but not necessarily sequential in time	Decision aids: expert system, human expert, human novice	1. Questionnaire 2. Interaction with computer Trust is easy to degrade and hard to reaffirm when predicted behavior does not occur
Muir (1987, 1994), Lee and Moray (1992, 1994), Muir and Moray (1989, 1996)	Trust in automation	1. Trust is a multi-dimensional construct 2. Supervisory control demands that the system be trustworthy 3. Transition of trust	1. $\text{Trust} = E[\text{natural and moral social persistence}] + E[\text{technical competent performance}] + E[\text{fiduciary responsibility}]$ where E means expectation 2. $\text{Trust} = \text{Predictability} + \text{Dependability} + \text{Faith} + \text{Competence} + \text{Reliability} + \text{Responsibility}$	Use of automation: manual/automatic	Process control simulation (Pasteurizer plant) 1. Although trust will fail sometimes when automation fails, it will recover again after a period of time 2. High positive correlation between operators' trust in and use of automation 3. Inverse relationship between trust and monitoring of the automation

Table B.2 Human Factors Notions of Trust, cont'd.

Literature	Problem domain	Definition	Basic model	Parameters	Measurements/Findings
Sheridan (1988)	<ol style="list-style-type: none"> 1. Concepts of rational/irrational trust 2. Cause and effect 	<p>Seven attributes:</p> <ol style="list-style-type: none"> 1. Reliability 2. Robustness 3. Familiarity 4. Understandability 5. Explication of intention 6. Usefulness 7. Dependence 	<p>Seven quantitative models:</p> <ol style="list-style-type: none"> 1. Bayesian expectation 2. Subjective expected utility 3. ROC model 4. Calibratability model 5. Direct subjective judgement of trust 6. Understandability models 7. Intention models 	Attributes of trust the author explored	
Singh, Molloy, Parasuraman (1992, 1993a, 1993b)	Attitudes toward automation technology that reflect a potential for complacency in aviation	<ol style="list-style-type: none"> 1. A psychological state characterized by a low index of suspicion 2. Accident sense as the cultivated capacity to foresee and forestall the development of situations conducive to critical malfunction and oversights 	Aviation Safety Reporting System	<ol style="list-style-type: none"> 1. General complacency 2. Confidence-related complacency 3. Reliance-related complacency 4. Trust-related complacency 5. Safety-related complacency 	Complacency Potential Rating Scale
Wickens (1994)	<ol style="list-style-type: none"> 1. Flight deck automation 2. Workload 3. Situational awareness 4. Perceived control 	<p>Overtrust: may result from a failure to initially calibrate a level of less than perfect automation. If one believes initially that the automation may be perfect, then minimal monitoring or 'second guessing' will be the consequence</p> <p>Mistrust: once trust is lost in automation because the latter fails to perform as expected, it may be hard to restore, and this brings us to the issue of mistrust</p>	<p>Sources of mistrust:</p> <ol style="list-style-type: none"> 1. Failure to understand (poor mental model) 2. Automation failure 3. Perceived automation failure 	Human-centered automation	<p>Some aspects should be implemented:</p> <ol style="list-style-type: none"> 1. Automation control and simplification 2. Interface/Display design 3. Training 4. Corporate policy

REFERENCES

- Bliss, J. P. (1997). Alarm reaction patterns by pilots as a function of reaction modality. *The International Journal of Aviation Psychology*, 7(1), 1-14.
- Chambers, A. B. & Nagel, D. C. (1985). Pilots of the future: Human or computer? *Communications of the Association for Computing Machinery*, 28(11), 1187-1199.
- Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution*, 2, 265-279.
- Gill, J. & Butler, R. (1996). Cycles of trust and distrust in joint ventures. *European Management Journal*, 14(1), 81-89.
- Hamsher, J. H., Geller, J. D. & Rotter, J. B. (1968). Interpersonal trust, internal-external control and the Warren Commission Report. *Journal of Personality and Social Psychology*, 9, 210-215.
- Knapp, R. K. & Vardaman, J. J. (1991). Response to an automated function cue: An operational measure of complacency. *Proceedings of the Human Factors Society 35th Annual Meeting*, 112-115.
- Larzelere, R. E. & Huston, T. L. (1980). The Dyadic trust scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and the Family*, 42(3), 595-604.
- Lee, J. D. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243-1270.
- Lee, J. D. & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153-184.
- Lerch, F. J. & Prietula, M. J. (1989). How do we trust machine advice? In G. Salvendy and M. J. Smith (Eds.), *Designing and using human-computer interface and knowledge based systems*. Amsterdam: Elsevier Science Publishers.
- Muir, B. M. (1987). Trust between human and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27, 527-539.
- Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automatic systems. *Ergonomics*, 37(11), 1905-1922.

- Muir, B. M. & Moray, N. (1989). Operators' trust in and use of automatic controllers. *Proceedings of the Annual Conference of the Human Factors Association of Canada*, 163-166.
- Muir, B. M. & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429-460.
- Rempel, J. K., Holmes, J. G. & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, 49(1), 95-112.
- Rempel, J. K. & Holmes, J. G. (1986). How do I trust thee? *Psychology Today*, 20(2), 28-34.
- Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, 35, 651-665.
- Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist*, 26(5), 443-452.
- Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, 35(1), 1-7.
- Schlenker, B. R., Helm, B. & Tedeschi, J. T. (1973). The effects of personality and situational variables on behavioral trust. *Journal of Personality and Social Psychology*, 25(3), 419-427.
- Sheridan, T. B. (1988). Trustworthiness of command and control systems. *IFAC Man-Machine Systems*, 427-431.
- Singh, I. L., Molloy, R. & Parasuraman, R. (1992). Development and validation of a scale of automation-induced "complacency." *Proceedings of the Human Factors Society 36th Annual Meeting*, 22-25.
- Singh, I. L., Molloy, R. & Parasuraman, R. (1993a). Automation-induced "complacency": Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology*, 3(2), 111-122.
- Singh, I. L., Molloy, R. & Parasuraman, R. (1993b). Individual differences in monitoring failures of automation. *The Journal of General Psychology*, 120(3), 357-373.
- Strickland, L. H. (1958). Surveillance and trust. *Journal of Personality*, 26, 200-215.

Wickens, C. D. (1994). Designing for situation awareness and trust in automation. *IFAC Integrated Systems Engineering*, 365-370.

Wrightsman, L. S. (1966). Personality and attitudinal correlates of trusting and trustworthy behaviors in a two-person game. *Journal of Personality and Social Psychology*, 4, 328-332.

APPENDIX C

Seong, Y. and Bisantz, A. M. (1998).

**Modeling Human Trust in Complex,
Automated Systems.**

**Poster presented at the Third
Automation Technology and Human Performance Conference,**

March 25-28,

Norfolk, VA. Old Dominion University.

Modeling Human Trust in Complex, Automated Systems Using a Lens Model Approach

Younho Seong, Ann M. Bisantz
State University of New York at Buffalo

INTRODUCTION

Automation has played an important role in supporting human and system performance in complex modern systems, such as aviation and process control. The advent of automation has changed the role of the human operator from performing direct manual control to the management of different levels of computer control. Human operators assume roles as supervisory controllers, interacting with the system through different levels of manual and automatic control (Sheridan & Johanssen, 1976). Therefore, the human operator must understand how to interact with the automated system, how the automation works, how to respond to system outputs, and how and when to intervene in the process, if the process fails. One factor affecting this interaction is the operator's trust in the automated system. Sheridan (1980) emphasizes the importance of human trust in automation as playing a key role in determining the level of a human operator's reliance on and the degree of intervention in automation and appropriate use of automation. Trust has been studied mainly from a sociological perspective which focused on interpersonal relationship between individuals. Following the sociological definitions of trust, more recent studies (Muir & Moray, 1996; Lee & Moray, 1992) have constructed models of human operator's trust in automated systems and shown how human trust in automated process control systems may affect system performance. These studies have focused on determining the extent to which human operator's trust in machines might affect system performance, and if so, identifying potential factors affecting the level of the operator's trust. An important concept regarding human trust is the notion of calibration: operators must have an appropriate level of trust in the information or automated system, given the characteristics of the situation. As Muir (1994) indicated, "well-calibrated" operators are better able to utilize automated systems. In case of aided adversarial decision making systems, understanding how well an operator judges the level of data integrity, based on the observable characteristics of the situations, becomes a very critical issue. We propose that Brunswik's Lens Model of human judgments (Brunswik, 1955; Hammond, Stewart, Brehmer & Steinmann, 1975) may be useful in formalizing the study of trust. The Lens Model provides dual models of a human judge and the environment to be judged, and allow the extent to which an individual's judgment behavior captures the structure of the environment to be assessed. This extent can provide a description of how well an operator's trust in the information, is calibrated to the actual environmental situation, as described by the relationship between those characteristics and the actual integrity of the information.

PREVIOUS MODELS OF HUMAN TRUST

Sociological Models of Human Trust

Rempel, Holmes, and Zanna's (1985) definition of trust contains critical aspects of trust which can be used to examine human trust in automation from the human factors perspective.

They emphasized not only components of interpersonal trust, but also the dynamic characteristics of trust toward a partner, regarding trust as a generalized expectation related to the subjective probability which an individual assigns to the occurrence of some set of future events (Rempel, et al., 1985). That is, the study suggested that humans evaluated their partners based on the characteristics they observed. Therefore, these characteristics served as cues to determine the level of human trust. The study is valuable in that it allows us to understand the importance of the role of human trust in the sociological domain and to identify certain characteristics of trust. Other research has also suggested that trust is a multi-factorial concept (Barber, 1983; Zuboff, 1988).

Human Factors Models of Human Trust

Based on Barber's (1983) study of trust, Muir (1994) constructed a hypothetical model of trust in machines, consisting of a linear combination of the characteristics identified by Barber. That is, the model represents human trust as a combination of persistence, technically competent performance, fiduciary responsibility, and interaction effects between these characteristics. In addition, Muir produced an integrated framework or model by crossing Barber's (1983) dimensions of trust, with Rempel et al.'s (1985) framework of trust as a process of hierarchical stages, developing over time. Lee & Moray (1992) extended the Muir's work and established a dynamic, mathematical model of trust based on a series of experiments. The model reflected dynamic characteristics, in that the current level of trust was affected by the previous level of trust and system-oriented factors such as the presence of automation faults and level of joint system performance. Both models, however, failed to explicitly consider the calibration of trust and the true state of automation trustworthiness, although the need to assist human operators in calibrating their trust was recognized. We propose that Brunswik's Lens Model may be valuable in describing human trust in automated systems, since it provides a mechanism for capturing this notion of calibration as well as the true state of automation trustworthiness.

LENS MODEL

Brunswik's Lens model, shown in Figure C.1, is a symmetrical framework in which describes how both the *environmental structure* and patterns of *cue utilization* collectively contribute to judgment performance. In this model, the judge combines cue information (X_i) about the environment to make a judgment (Y_j). The model represents the classical notion of information transformation from stimulus (information presentation) to response (judgment) in which humans process information internally to yield some functional response based on cues observed, which in turn are representations of the environmental state. Thus, the model includes not only a classical decision concept, i.e., how humans sample and combine cues presented to them, but also the relationship between available cues and the true state of the environment. By analyzing a judge's cue utilization policy, therefore, we may be able to understand how that judge has adapted to the structure of the environment. The predictability of the environment, given a set of cues (the ecological validity of the cues) can also be assessed. Therefore, this model allows us to assess and evaluate how well the true environment structure is represented via a set of cues. Additionally, achievement, denoted as r_a , represents how well human judgments correspond to the actual values of the environmental criterion to be judged. Achievement is

shown in Figure C.1 as a line connecting judgments to criterion values. Because the Lens Model provides the means for considering the judge's adaptation to the environment, and the degree of achievement, both of which relate to the calibration of human trust, it seems that the use of the Lens Model approach to model human trust in automated systems is reasonable.

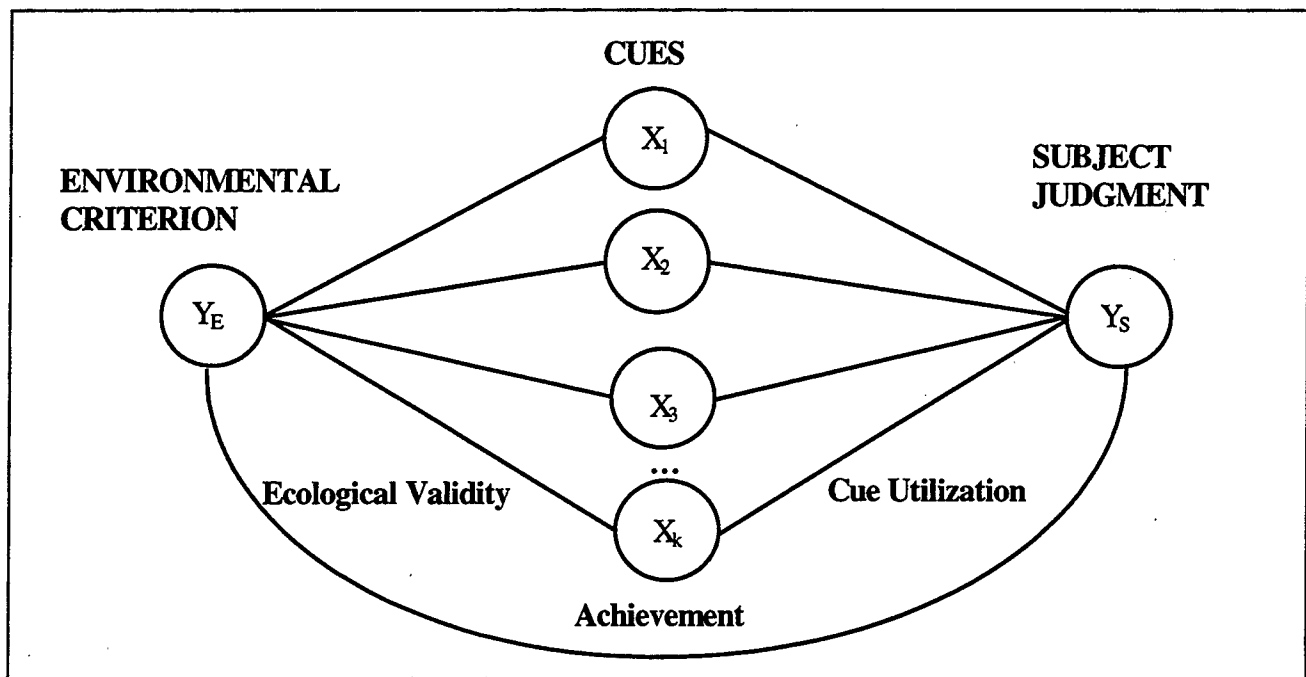


Figure C.1 Brunswik's Lens Model

APPLYING THE LENS MODEL INTO HUMAN TRUST IN COMPLEX, AUTOMATED SYSTEMS

Conceptually, modeling human trust in automated systems using the Lens Model is relatively straightforward. The judgment modeled in this case is the operator's judgment of the trustworthiness of some system component or output. That is, the operator decides whether or not a system component is to be trusted. In Lens Model terms, then, the environmental criterion is the actual trustworthiness of the component. The judgment is the operators' assessment of that trustworthiness. To make this judgment, the operator must rely on a set of observable cues which have some relationship to the components trustworthiness. In this paradigm, the concept of calibration is explicitly measured by achievement (r_a)—the extent to which the operator's assessment of trustworthiness matches the true state of the environment. One can also consider calibration to include operator's adaptation to the structure of the environment, in terms of the relationship between the cues and actual integrity of information.

Further specification and experimental verification of this model of trust in automation beyond the general level noted above presents certain difficulties, however. First, there is no clear, objective measurement of the true state of environment, in terms of its trustworthiness.

Generally, trust as a state in itself has been measured only subjectively. This is problematic in terms of the Lens Model formulation, since application of the Lens Model and evaluation of the model parameters requires knowledge of the *true* environmental state. To circumvent this difficulty, we propose transforming the judgment from one of an assessment of trustworthiness to one that is more performance oriented. From an engineering standpoint, we are interested in human trust in a system to the extent to which that trust affects system performance. For instance, we are interested in whether or not operators utilize an automated controller, or obtain certain data, given their trust in that controller or information source. The true state of the environment, in terms of the adequacy of the controller, or the integrity of the data source, can be objectively determined. For these examples, the operator judgment would be whether to use the controller or the data. More generally, the operator judgment is one of component utilization, and true state of the environmental criterion is whether or not the component should have been used. In terms of trust, this assumes that an operator's behavior in utilizing a system component reflects their trust in that component.

Second, to implement a Lens Model description of human trust in automation, it is necessary to specify what cues might be available for an operator to make a judgment about whether to use a system component. Candidate cues include the components of trust identified by previous studies of trust (e.g., Barber, 1983; Rempel, et al., 1985; Zuboff, 1988). For instance, cues could include such factors as predictability, dependability, faith, reliability, competence or robustness. To be included in a quantitative Lens Model, these cues would be both measurable, and available to the operator. The availability of these candidate cues to the operator depends to some extent on how information is displayed to operators. However, the consideration of how to measure these cues must be addressed. For example, consider predictability. If we define the environment to be judged in terms of a subsystem, or set of systems, we can represent predictability in terms of the degrees of freedom in performance that were designed into the system. That is, predictability could be measured in terms of allowed error or performance variance. The smaller the degree of freedom, or allowable error, the more predictable the system is. If predictability is one component of trust, as Barber claimed, then trust will be negatively impacted by a large degree of performance variability. Additionally, the reliability of a system or component could be measured in terms of past performance (e.g., breakdowns, errors, etc.).

Instantiating the Model

To evaluate the model, an experimental framework has been established in an Information Warfare (IW) domain (Seong, Llinas, Drury, & Bisantz, 1998) in which one can consider trust in the context of aided adversarial decision making, where military officers must assess the integrity of information which may be intentionally altered or degraded by an enemy. In this domain, the points of attack by an enemy can be the real battle situation, data gathering or fusion algorithms, or a data transfer network. By changing the points of simulated attack, we may be able to observe how operators successfully calibrate their trust in terms of accurately pinpointing the point of attack, and changing the level of trust. In the IW domain, studying human trust is important for several reasons. For example, forces might be vulnerable to information attacks which diminish their trust in data fusion or other decision aids, rendering these assets less useful, or to deceptive attacks, in which an inappropriately high level of trust in the aid is maintained. In terms of the Lens Model approach, data, fusion algorithm outputs,

would be judged as usable or not (e.g., trustworthy or not), based on operators understanding of the predictability, reliability, etc. of the information displayed to them.

SUMMARY AND FUTURE RESEARCH

A Lens Model approach for modeling human trust in automated systems has been proposed (Figure C.2). Because the Lens Model provides the means for modeling both human judgment policy and the actual structure of the environment, it allows operator calibration to the actual trustworthiness of a system to be explicitly considered. Conceptual solutions for addressing certain difficulties with this approach, such as the objective determination of the true state of system trustworthiness, and the identification and measure of cues which reflect system trustworthiness, were discussed. Finally, an experimental framework in the domain of Information Warfare was described, which may provide the means for further instantiating and evaluating the effectiveness of this model of human trust in automation.

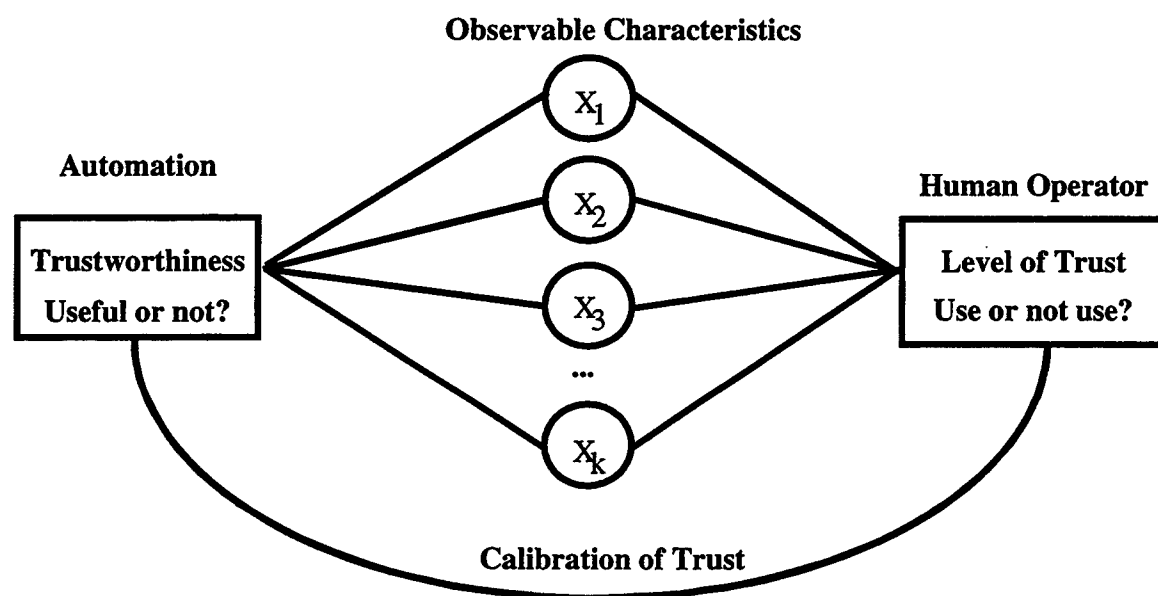


Figure C.2 Model of human trust in automation using the Lens model

REFERENCES

- Barber, B. (1983). *The Logic and Limits of Trust*. New Brunswick, NJ: Rutgers University Press.
- Brunswik, E. (1952). *The Conceptual Framework of Psychology*. Chicago, IL: University of Chicago Press.
- Cooksey, R. W. (1996). *Judgment Analysis: Theory, Methods and Applications*. New York: Academic Press.
- Hammond, K. R., Stewart, T. R., Brehmer, B. & Steinmann, D. O. (1975). Social judgment theory. In M. F. Kaplan and S. Schwartz (Eds.), *Human Judgment and Decision Processes*. New York: Academic Press.
- Lee, J. D. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243-1270.
- Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11), 1905-1922.
- Muir, B. M. & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429-460.
- Rempel, J. K., Holmes, J. G. & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, 49(1), 95-112.
- Seong, Y., Llinas, J., Drury, C. G. & Bisantz, A. M. (1998). Human trust in aided adversarial decision-making systems. In M. W. Scerbo (Ed.), *Automation Technology and Human Performance: Current Research and Trends*. Mahwah, NJ: Lawrence Erlbaum.
- Sheridan, T. B. (1980, October). Computer control and human alienation. *Technology Review*, 61-73.
- Sheridan, T. B. & Johannesssen, G. (1976). *Monitoring behavior and supervisory control*. New York: Plenum.
- Zuboff, S. (1988). *In the age of the smart machine: The Future of work and power*. New York: Basic Books.

APPENDIX D

Seong, Y., Llinas, J., Drury, C. G., and Bisantz, A. M. (1998).

Human Trust in Aided Adversarial Decision-Making Systems.

Poster presented at the Third Automation Technology and Human Performance Conference,

March 25-28, Norfolk, VA.

Old Dominion University.

Human Trust in Aided Adversarial Decision-Making Systems

Younho Seong, James Llinas, Colin G. Drury, Ann M. Bisantz
Center for Multisource Information Fusion
State University of New York at Buffalo

INTRODUCTION

The study of trust has a long history in sociological literature; however, that history is not rich in empirical studies. Some representative, recent studies of human trust in automation have been performed on a continuous chemical process control simulation (Muir & Moray, 1996; Lee & Moray, 1992, 1994). This means that the objects which human operators dealt with were machines or displays representing the behavior of machines. Thus, the induced characteristics of human trust in automation were essentially concerned with predictability, dependability, and faith, whose attributes could be easily captured from the behavior of machines. Current automation, which in the context of this study is technology using a data fusion process, produces estimates of situational conditions which are, ideally, of reasonable but imperfect quality, i.e., there is some uncertainty in the estimates. However, this study extends the studies of trust to situations where, additionally, the automation system is open to deliberate manipulation by adversaries. Having another person or a group of people *at the other end* changes the central scheme of the problem. In this environment, called *Information Warfare* (IW), human operators must deal with faults resulting from both mechanistic automation failure (imperfect data fusion) and premeditated deception or misguidance manipulated by an adversary. Human operators should have the ability to distinguish the faults perpetrated by the foe, to calibrate their trust in decision-making aids and to eventually accomplish their mission efficiently. Without knowing the system's vulnerability from an adversary, human operators may regard faults of corrupted information by an adversary as automation failures in the data fusion process or malfunction of displays. In this paradigm, some characteristics of trust, which were identified in previous studies, such as fiduciary responsibility, may not be applicable because of the hostile environment. Simultaneously, the new paradigm should be able to include the simple relationship between human operators and decision-making aids.

POTENTIAL CHARACTERISTICS OF HUMAN TRUST AND IMPLICATIONS FOR IW

Among the sociological studies defining trust in the interpersonal relationships, Rempel, Holmes, & Zanna (1985) characterized trust as a multi-faceted construct having three dimensions. While this represents one classification of trust characteristics, Sheridan (1980) suggested a more comprehensive set of seven possible characteristics of human trust in the human-machine systems. As we are dealing in IW with trust of machines, and the data fusion processes, the Sheridan's classification is a better starting point. We will consider the applicability of Sheridan's characteristics in turn where the domain is IW.

The first of Sheridan's aspect of trust in automation seems to relate to notions of *reliability*. This implies a system of reliable, predictable, and consistent functioning. In other words, a sense of reliability in a decision aid (DA) is established when it can be observed to create the same output repeatedly under a particular set of circumstances. There are three fundamental points of vulnerability from the real world situation to the display for operators. These are true input, observable input and displayed output. True input is susceptible to deception and corruption of operations and actions by artificially creating a false circumstance. Observable input is susceptible to parametric or algorithmic corruption. This is creating false output from internal processes in spite of a *correct* circumstance in the real world. Displayed output is susceptible to algorithmic corruption. This is creating a false display in spite of both (a) an unperturbed real world and (b) unperturbed (non-display related) processing operations.

The second aspect of trust is a sense of *competence* or *robustness* in the DA. That is, robustness supports expectations of future performance based on capabilities and knowledge not strictly associated with specific circumstances that have occurred before. In some sense, this represents a *calibration* of the decision-aiding software processes or algorithms. Corruption of a operator's trust in the DA through IW techniques places a requirement on the deceiver to know the nature of adaptivity in the DA internal workings. However, if the operator is astute in the likely technology for the DA, a reasonable approximation to the inner workings may be feasible. Alternately, if the deceiver is aware of the limits of capability of the DA, the operator could create circumstances which are in those fringe operating regions and reduce the operators impressions of the competence of the DA.

Familiarity is the third feature or aspect of trust. Often a person confronts a situation or an object with a high degree of novelty, but still feels familiar with and able to deal with the situation. Either from a naturalistic or inherent cultural expectation, familiarity may prevent any exploratory risk-taking behavior to diagnose the situations, or to identify objects whether new or familiar. Consequently, it may induce biased decision-making. Because of the fact that familiarity is not based on any scientific knowledge or expertise and tends to be inherited from those who have cultural similarity with us, the person who is confronting an unfamiliar or unanticipated situation or object will be very vulnerable to deception. Unlike other industrial settings where unanticipated, and so unfamiliar events are sometimes confronted by operators, operators in military command, control, communication and information (C³I) systems may not have been exposed to or trained in unanticipated events. For automated DA's, this can involve the use of common or perhaps military-standard symbology, nomenclature, etc.

Sheridan's fourth characteristic is *understandability*. The construct of understandability is equivalent to developing an appropriate mental model, possibly with the aid of familiarity. In designing a machine to aid an operator, understandability usually is affected by the degree of transparency of the system which the operator can *see* through the interface to the underlying system. Opaque machines or interface media will not only prevent the operator from trusting the machines, but also from engaging in problem-solving activities in cases of warnings or mishaps. Thus, any means by which an adversary could corrupt the graphic user interface or other interface functions, in order to confound the operators' ability to understand a system, would lead to distrust. This is particularly important if the operators do not participate in the development process. If this is the case, then the degree of understandability governs the operators' ability to

form his own mental model of the inner workings of the DA, which also lends itself to predictability. Here the deceiver can simply have the goal of generating randomness in his or her attack, since irregular patterns between input and output will aid in loss of trust.

Next we consider *Explication of Intent*. Instead of leaving a person in a position where the covert meanings have to be discovered and understood from the systems' behavior, this attribute allows people to trust others over those who just perform tasks. However, current technological improvements in designing intelligent computers are not well enough developed to allow operators to communicate using higher level intentions. Unless we develop intelligent machines, which can specify their intentions of future actions outright, we have to rely on the current available technologies, e.g., in the form of symbols, short statements, or a combination of both which are pre-programmed by system designers. Therefore, we are often forced to trust (or not to trust) based on a symbolic medium through which one produces effects and on the basis of which one derives an interpretation of "what is happening." Estimating intent is one of the more difficult things to do in military situations but is also, if done correctly, one of the highest payoff areas. Intent is approximately an explicit indication of a future planned action to which the actor is committed; the difficulty is in defining and observing those indicators. An actor develops intent based in part on his value system; that is, the operator will plan actions that have a sense of payoff in the context of a value system. The value system is, however, multi-dimensional; it has relatedness to military goals and objectives but also to notions of personal value and the societal notions of value we all grow up with; these are the harder elements to judge. The development of an intended action is dependent on a particular or sequence of particular outputs (from a DA) in the context of judged value;

$Action(T + \Delta T) \approx (DA\ Output \mid T, DA\ Output \mid (T-1), \dots \text{and implied/calculated value of action, given the DA Output})$

By creating any false output of the DA (by any means), the deceiver corrupts this relationship and leads the deceived toward taking an alternative action.

Usefulness is Sheridan's sixth trait associated with trust. Usefulness of data or machines means responding in a useful way to create something of value for operators, eventually developing into trust. In fact, one branch of decision theory is explicitly based on such values: "Utility theory." This, however, raises a question: Does usefulness of data ensure the quality of decision-making, or make operators dependent on the DAs? In other words, do notions of data values help decision performance, induce trust, or both? Studies indicated that humans tend to behave in different ways rather than using the estimated utility (e.g., Tversky & Kahneman, 1974; Pulford & Colman, 1996; Klein, 1997). Usefulness relates to the ability of a DA to respond in a useful or responsible way and to notions of value and utility of the DA itself in a local or specific sense. By this latter remark, it is meant that if the DA performs and produces output that is consistent with its planned concept of employment, then it will be judged useful in this sense. However, Usefulness also takes us back to the notions of value; DA's can be useful but not valuable. Most deceivers would probably not have insight into notions of usefulness since this can mean many things, but they may have some notions of value which can be exploited as described above.

The final aspect of trust in Sheridan's classification is Dependency. Trust is not a useful concept unless an operator is willing to depend on a machine. Dependency is one aspect of trust accessible to empirical measurement, i.e., by the fraction of time an operator behaves as if the machine were trustworthy. From Wickens (1992), we see that likelihood of use of automation is commensurate with level of trust in automation and inversely related with level of self-confidence of operator which implies a number of things. For instance, if the operators are known to be poorly trained then their self-confidence would presumably be low and their reliance on automated DA's high, perhaps to an unwarranted degree. Also, if the DA has only been built to handle easy cases and has never confronted the fringe-condition where it can fail, the operator may develop an unwarranted trust in its competence.

STRUCTURING EXPERIMENTS FOR INVESTIGATING TRUST IN AN IW DOMAIN

Given the tools for measuring trust that were presented in human factors literature, it is possible to consider the types of empirical studies which could be performed to investigate human trust in IW domains. Such controlled investigations would provide a better understanding of what situation characteristics influence trust, as measured by the either or both of the psychophysical ratings and performance and process measures, and also how changes in an operator's trust in system components affect ultimate system performance. In order to develop possible scenarios for investigating aspects of trust in IW, it is instructive to consider how trust has been investigated in other automated systems.

Previous Investigation of Trust in Automated Support Systems

As described before, empirical work in the area of human trust in an automated support (decision-aided) system is limited, and has concentrated primarily on investigating trust in simulated, semi-automated process control environments. Moreover, and importantly as regards our concerns for IW environments, these studies have been in *non-adversarial* domains (i.e., Muir & Moray, 1996; Lee & Moray, 1994). Different system aspects were altered to see how participants' trust in systems components, such as the automated controller, was affected. In particular, Muir & Moray (1996) altered the quality of the pump systems by introducing either *random* or *constant errors* in its ability to maintain a set-point, introduced errors into the pump's *display* of its pump rate (although actual pump rate was error-free), and the performance of the automated controller in setting and maintaining appropriate settings for the pump. Lee and Moray introduced *faults into pump performance* (Lee & Moray, 1992) or *faults into either automatic or manual controllers* (Lee & Moray, 1994). These conditions are not unlike the type conditions that may arise in IW environments. Trust was measured both subjectively, using rating scales, and objectively, by logging participants' actions (e.g., hypothesizing that more or less use of an automated control system implied more or less trust in that automated system). Because faults were introduced into different components, these experiments investigated trust in a particular system aspect rather than trust in automation generally.

Designing Experimental Scenarios for Studies of Trust in AADM Environments

In the AADM environments of interest to this study, pictured in Figure D.2, data fusion techniques are used to aid the decision-maker by synthesizing data from numerous sources into a form useful for the decision-maker. Because the environments of interest are ones involving adversaries, the possibility of corruption in either or all of the data, fusion algorithms, and displays involved in such decision-aiding systems can be introduced by *Information Operations* manipulated by the hostile forces. It is in such environments that we would like to perform human-in-the-loop experiments to study various hypotheses related to human trust under IW conditions and in AADM environments. As mentioned before, the literature is not very helpful in this regard. The multi-faceted manner in which trust was investigated in the above experiments suggests two dimensions along which studies of human trust in complex environments, such as an AADM environment, could vary: we called these the *system dimension*, and the *surface-depth level*.

System Dimension. In the pasteurization experiments (Lee & Moray 1994; Muir & Moray, 1996), the quality of system performance was manipulated at what could be called different *system* levels. Faults or random errors were introduced at the *level of the (system) environment*; the process control system itself (i.e., the pumps), and at the *level of a (system) control intervention* (i.e., the automated controller). There are analogous levels in an AADM environment. The physical component level in the pasteurization experiments—the pumps and heaters—corresponds to the *actual tactical situation* that is taking place. Just as the states of pumps and heaters can be observed and controlled, the states of hostile and friendly assets can be assessed, and actions related to the situation can be taken. The next level, *data fusion systems and algorithms*, which automatically combine and synthesize information obtained from the tactical environment, can be considered analogous to the automated controller in the pasteurization experiments, which used information from the physical control system to automatically take control actions. Finally, in an AADM environment, one can consider a third level, the *interface level*. At this level, the results of the data fusion algorithms are displayed to the operator, in order to aid decision making.

Surface-Depth Level. Another dimension along which investigations of trust can vary is a *surface-depth level*. The *surface level* corresponds to the information available about the environment (as formalized in Brunswik's Lens Model; Cooksey, 1996), whereas the *depth level* corresponds to the actual state of the environment. The manipulations performed by Muir & Moray (1996) can be described in terms of these dimensions. Muir & Moray (1996) manipulated both the characteristics of the pump itself (depth level) and the display of the pump rate (surface level). This surface-depth dimension can be applied at all three of the system dimension levels described above, resulting in six combinations (see Figure D.1).

Further Categories of Corruption. Within each cell of upper portion of Figure D.1, it is possible to identify various types of malfunction, or causes of information degradation or corruption.

- **Degradation.** The *quality* of the system component can be degraded through constant, random errors, or discrete failures.

- Failure. System components can *fail* completely resulting in a loss of data.

Different causal factors for the corrupting processes can also be considered:

- Non-intentional. System components can degrade due to non-intentional malfunction.
- Sabotage. An enemy can take *intentional action* to interfere with a system component.
- Subterfuge. An enemy can take intentional action both to interfere with a system component, *and to disguise that sabotage*.

Given a particular experimental context, the surface-depth and system dimensions, along with the levels of malfunction, and causal factors, can be used to systematically define a series of experimental manipulations which can be used to investigate issues of trust in AADM environments (Figure D.1).

Design of Experimentation. For studies of AADM, a possible experimental context would be an interactive battle simulation in which people must make interpretations and/or decisions (e.g., identification of unknowns, decisions to engage hostile forces) based on information gathered and fused into decision-aiding estimates about the situation, such as those related to electronic emissions, weapons profiles, and locations and movements of various agents. The simulation would include data fusion modules which could synthesize environmental information in order to aid the participants decisions (Figure D.2, on the next page).

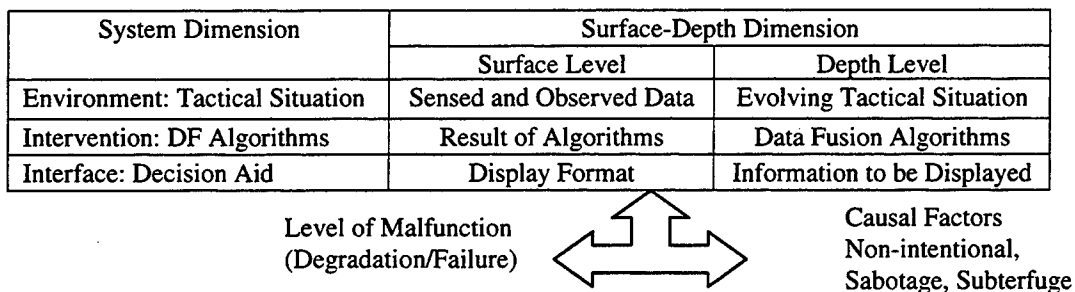


Figure D.1 Components of an AADM Environment described along a system and a surface-depth dimension. Potential experimental scenarios and manipulations, organized by system and surface-depth dimensions and levels of malfunction, causal factors.

In either cases of having single or multiple participants, it is also possible to consider that some agents or participants would be synthetic; that is, created in software as so-called *intelligent agents*. This would add a dimension of experimental control, since the behavior of the agent would be fully controllable, or at least controllable within known limits. If one is to study such environments experimentally, a synthetic environment of this type needs to be created in a controlled way. In developing such simulation environments, one of the primary capabilities to establish is that of the *problem space* or the framework from which problem information, data, and parameters evolve. Some call this type of capability a *scenario generator*, in which simulated observable cues are produced as inputs to processes under study. In our case, we are proposing the use of an existing capability for this, the so-called “SAGE” (Semi-Automated Ground

Environment) software system. SAGE can also provide representation of two adversarial commanders, or at least the information environments in which their decision-making is being conducted.

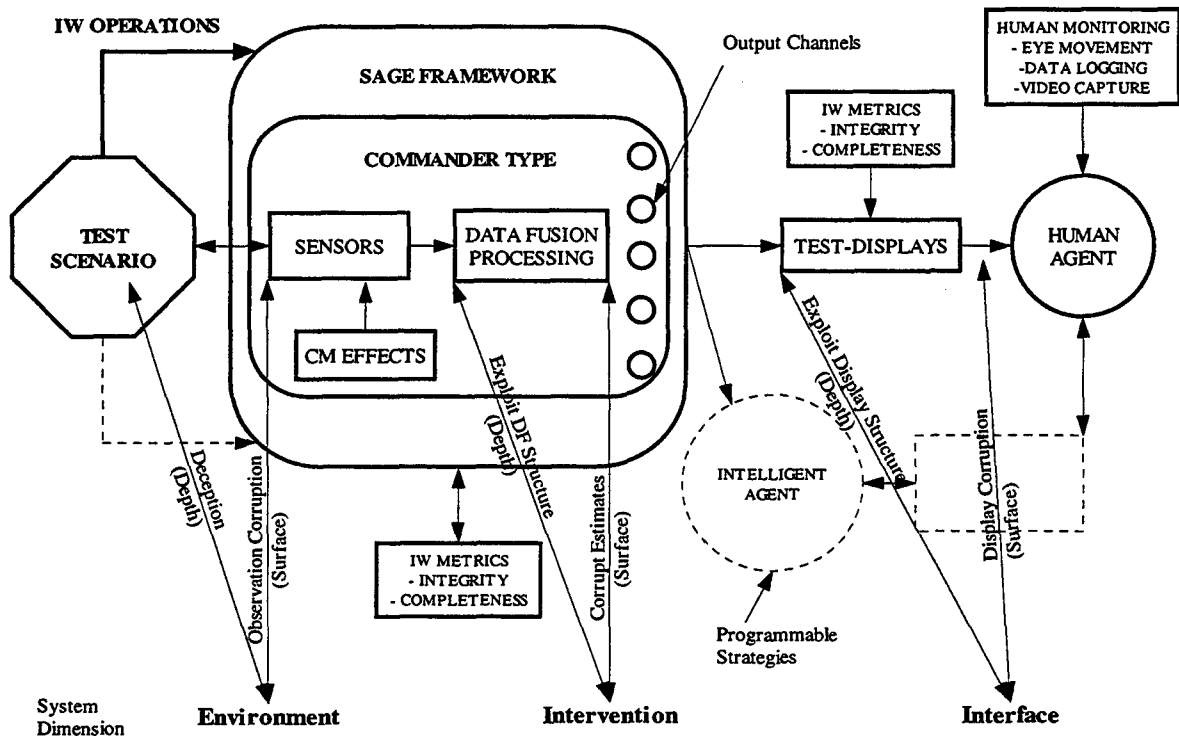


Figure D.2 Experimental concept for AADM IW

SUMMARY AND FUTURE RESEARCH

Applicability of the Sheridan's classification into AADM systems has been investigated. Because the classification offers a variety of characteristics of trust in supervisory control systems, it forms a good Basis for the conceptualization of trust in AADM environment and further development of experimental framework. Due to the particular characteristic of the domain which adversaries, and multiple processes are involved, however, the experimental framework should consider the possible points of attack which could be varied from the real world to human operators. Based on the interpretation of the classification from the AADM perspective, then, three factors, and scenarios are discussed for an experimental framework in AADM environment. These factors include the dimension of the point of attack, degree of attack, and position of attack which may affect human operator's role of monitoring activities and decision-making actions, which consequently may have an impact on operators' calibration of trust.

REFERENCES

- Cooksey, R. W. (1996). *Judgment analysis: Theory, methods and applications*. New York: Academic Press.
- Klein, G. (1997). The Recognition-Primed Decision (RPD) model: Looking back, looking forward. In C. E. Zsombok and G. Klein (Eds.), *Naturalistic decision making*. (pp. 285-292). Mahwah, NJ: Lawrence Erlbaum.
- Lee, J. D. & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243-1270.
- Lee, J. D. & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153-184.
- Muir, B. M. & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429-460.
- Pulford, B. D. & Colman, A. M. (1996). Overconfidence, base rates and outcome positivity/negativity of predicted events. *British Journal of Psychology*, 87(3), 431-445.
- Sheridan, T. B. (1980, October). Computer control and human alienation. *Technology Review*, 61-73.
- Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.
- Wickens, C. (1992). *Engineering psychology and human performance*, 2nd ed. New York: Harper-Collins.